AMSE Association for the Advancement of Modelling & Simulation Techniques in Enterprises
IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Modelling, Measurement and Control C
Vol. 82, No. 1-4, December, 2021, pp. 1-5
Journal homepage: http://iieta.org/journals/mmc_c

Object Detection, Localization and Tracking-Based Deep Learning for Smart Wheelchair

Louis Lecrosnier[1,2*], Redouane Khemmar[1,2], Nicolas Ragot[1,2], Romain Rossi[1,2], Jean-Yves Ertaud[1,2], Benoit Decoux[1,2], Yohan Dupuis[1,2]

[1] UNIRouen, Normandy University, ESIGELEC/IRSEEM, Saint-Etienne-du-Rouvray 76800, France
[2] Technopôle du Madrillet, Avenue Galilée - BP 10024, Saint-Etienne-du-Rouvray 76801, France

Corresponding Author Email: lecrosnier@esigelec.fr

https://doi.org/10.18280/mmc_c.821-401

Received: 12 March 2021
Accepted: 23 July 2021

Keywords:
deep learning, object detection, localization, tracking, distance estimation, mobile robotics, environment perception, smart mobility

ABSTRACT

In this article, we present our work regarding the development of advanced driver-assistance systems for an electric-powered wheelchair. Our project aims at improving the autonomy of people with reduced mobility. After conducting a clinical study, we identified several use-cases. In this paper, we introduce the detection, localization and tracking of points of interest in the immediate surroundings of the chair in an indoor environment, i.e.: doors, handles, light switches, etc. The aim is not only to improve perception around the chair but also to enable semi-autonomous driving towards these targets. First, we introduce a repurposing of YOLOv3, the object detection algorithm, to our use case. Then, we show our use of the Intel Realsense camera for depth estimation. Finally, we describe our adaptation of the SORT algorithm to track 3D interest points. To validate our approach, we realized several experiments in a controlled indoor environment. The detection, distance estimation, and tracking pipeline is tested using our custom dataset. This includes corridors, doors, handles, and switches. One of the scenarios studied to validate the proposed platform includes not only the detection and tracking of objects but also the movement of the wheelchair towards one of these points of interest.

1. INTRODUCTION

Detection, classification, positioning, and tracking of objects are necessary tasks in the field of mobile robotics. Nowadays, these tasks often rely on computer vision and are carried using various sensor modalities (camera, LIDAR, RADAR) and processing algorithms (pattern recognition, filters, classification, feature extraction, etc.).

In this paper, we focused our work on the object detection, localization and tracking of interest points on an electric-wheelchair-type robotic platform.

Our contribution aims to develop object detection, depth estimation, and target tracking for the wheelchair, using deep learning approaches. We carried out the learning phases using a dataset we created, and then verified our pipeline within ESIGELEC's autonomous navigation platform. This article is organized as follows. The introduction is presented in section I. In section II, we present a state-of-the-art of object detection and tracking dealing with deep learning approaches. The architecture of the intelligent wheelchair is presented in section III. Our approach to object detection, depth estimation, and tracking is then detailed in section IV. Section V will conclude the article.

2. RELATED WORK

2.1 Object detection

Deep-learning methods are currently state-of-the-art for object detection tasks. Among these algorithms, methods like Fast R-CNN [1], Faster R-CNN [2] and Mask R-CNN [3] are generally composed of two stages of a convolutional neural network (CNN) [4] (a set of cells called neurons which, working together, independently make predictions from the network inputs). The first module provides Region of Interest (ROI) with attached coordinates, i.e. where an object (whatever its nature) would be showing in the image. This is the ROI proposal. The second module manages the object detection step. It provides a class prediction of the proposed region. The YOLOv3 (You Only Look Once) algorithm [5] is one of the most powerful deep learning object detection algorithms. YOLO is based on the principle of regression instead of classification. This means that the entire process of locating an ROI encompassing an object and classifying that object is done at once. This enables real-time object detection on a standard GPU. Similarly to YOLO, the SSD [6] algorithm combines methods for classifying and locating ROI. By this mean, SSD avoids the feature and pixels extracted for each bounding box from the image. In SSD, the VGG-16 [7] architecture is used for feature extraction for in the first layers of the neural network. The size of the next layers is progressively lowered to enable multi-scale detection (unlike YOLO). A set of detection prediction is generated by each of these layers.

2.2 Datasets

For the detection of objects, numerous datasets are freely available. We can mention some of them: ImageNet [8] with 14,000,000,000 instances per image. It includes 1000 classes [9], but the accessibility of this game of data remains limited.

SIFT10M [10] includes 11 million annotated images. It is structured in the form of reference points of known images (SIFT method). Open Images [11], which is a community dataset with more than 9000000 images. It comprises 7881 different classes [12], linked to labeled objects with ROIs. This dataset represents a wide range of classes, and is frequently expanded. COCO [13] contains 1.500.000 images. Objects are categorized and labeled. It includes 80 object classes [14]. COCO has been extended in COCO-stuff [15] and contains 1,800,000+ images with 181 classes [16]. PASCAL VOC [17] focuses on the detection of pedestrians and counts 500,000 labeled images with 20 classes [18]. CIFAR-100 [19] regroups more than 60000 images and includes 100 [20] different classes. INRIA [21] is another pedestrian detection database. So is also Caltech [22]. In general, these data sets are fairly rich in terms of the classes representation in urban environment (cars, bicycles, pedestrians, buses, etc.). The COCO dataset is one the reference in terms of object detection and is ideal for comparing different object detection models. The latter offers different databases adapted to the training phases and the inference of neural networks. Yolov3, for example, uses ImageNet to train the first 53 layers of its network and establish the reference databases. As this data set is very dense, it was a wise choice to prepare for Yolov3's training. This model then uses other databases for detection and classification, such as COCO for example.

3. ARCHITECTURE OF THE WHEELCHAIR-BASED PLATFORM

The robotic electric wheelchair of the IRSEEM laboratory is an Invacare, Bora model, which was stripped down from all the original electronics. We then added: 1. an embedded PC running Linux Ubuntu 16.04 LTS, 2. a Roboteq engine driver, 3. an Xbox controller, which can handle a USB or Bluetooth connection to the wheelchair computer, 4. a WIFI router to provide a wireless access point on the wheelchair, 5. an embedded HMI with a touch screen.

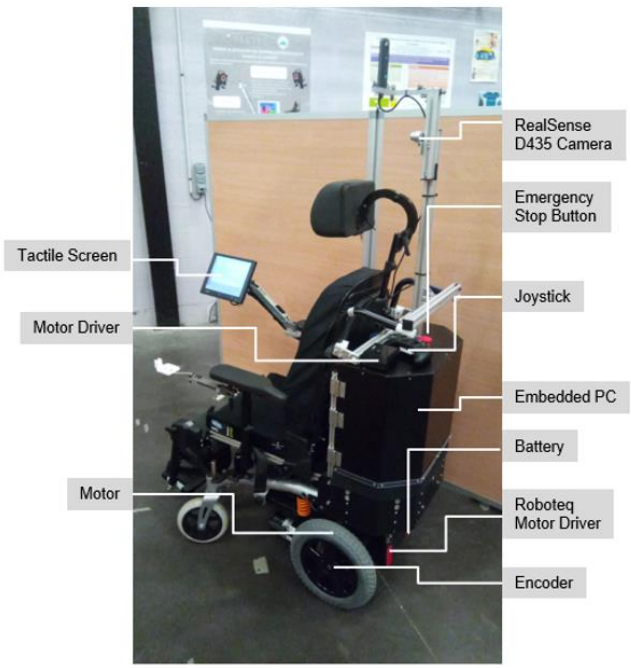

Figure 1. IRSEEM's electric and robotic wheelchair

All the software developments were carried out using the ROS robotic middleware, which seamlessly handles the multithreaded communication between the various software modules. The wheelchair hardware architecture is depicted in Figure 1. The on-board computed runs Ubuntu 16.04 OS, and works with a 250GB SSD and 8GB of RAM.

The wheelchair allows two ways of interaction. The added touch-screen is directly connected via HDMI to the embedded-computer, and provides a GUI with a visual feedback of the detected objects.

We rely on the Roboteq engine driver to control the motors. For connectivity purposes, the wheelchair can use its own Wi-Fi router. An Xbox One controller can also be used to control the wheelchair. We use Intel RealSense D435 camera to provide color and depth images to the perception software modules.

4. DEEP-LEARNING-BASED DETECTION, LOCALIZATION, AND TRACKING

4.1 Object detection

To carry out object detection, we based on the powerful neural network YOLOv3. In the ADAPT project framework, the detection and classification of indoor environment-specific features (such as doors, door handles, and switches) is an essential part of the essential point. Since these classes are under-represented in the YOLOv3 training dataset (i.e. ImageNet [23]), we composed a custom dataset that we used to perform network training. We extracted 755 door images from the MCIndoor20000 dataset [24], which consists of labeled images containing various indoor environment objects.

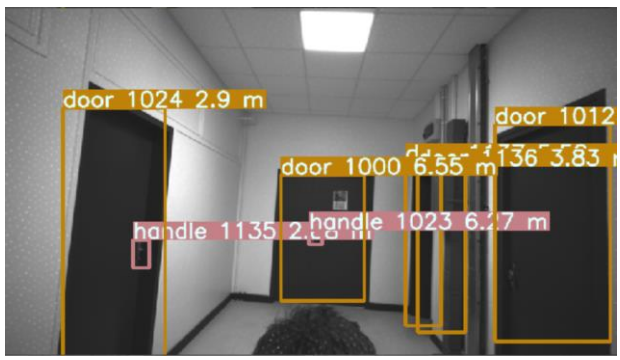

Figure 2. An example of doors and door handles object detection in an indoor environment

In the literature, it is very difficult to find an open dataset including sufficient representation of switches, segmented and labeled door handles. For this reason, we developed our custom dataset within ESIGELEC, consisting of 1885 images, which we combined with the door images from the MCIndoor20000 dataset. We supervised the labeling of 2640 images from two combined datasets by using a semi-automatic labeling tool we developed. Finally, we proceeded to re-train the YOLOv3 model for the required classes. For this transfer learning process, we trained only the classification layers of the neural network. The qualitative results of the detection process after re-training the YOLOv3 model on the recognition of doors, door handles, and light switches are shown in Figure 2.

4.2 Object tracking

Using the combination of the position and associated depth with each item detected in the scene by the object detection algorithm, we use the Simple Online and Real-time Tracking (SORT) algorithm [25, 26] to track the different objects in the scene while the wheelchair is in motion. Door handles, switches, and door objects are detected from a video stream and assigned a given distance.

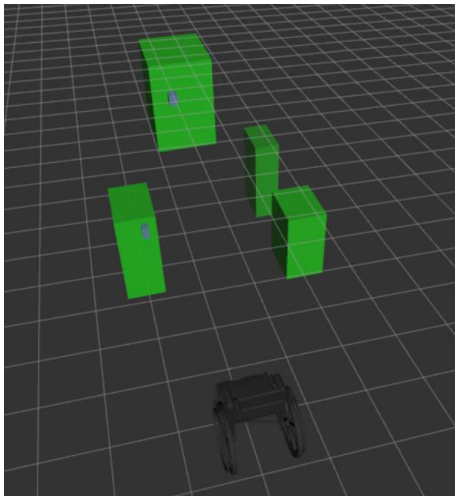

Figure 3. 3D local semantic map

SORT analyzes the detected objects and determines whether a given object is newly seen, or if the object's movement is a consequence of the wheelchair's movements. This Kalman filter-based algorithm finally provides a unique identification number to each newly detected object. SORT keeps track of multiple objects simultaneously and filters out the positions of noisy objects associated with moving boundary boxes.

Finally, we use the odometry data from the T265 RealSense camera to estimate the wheelchair's displacement. We combine this information with the object position to visualize a 3D semantic map of the environment. Figure 3 shows an example of a 3D semantic map containing detected and tracked objects in real-time.

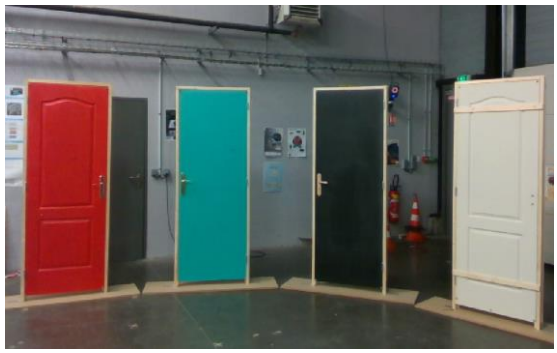

Figure 4. Arrangement of the doors in the validation dataset

5. EXPERIMENTAL VALIDATION

To validate our different developments, we have recorded an experimental dataset in the IRSEEM laboratory. The validation scenario includes an environment with four-doors having handles of different shapes and colors. The environment is instrumented with a Vicon motion capture system. The doors as well as the wheelchair (equipped with reflective markers necessary for localization) are localized by the Vicon motion capture system, which provides their position and orientation with millimeter accuracy at a frequency of 100Hz [27].

Table 1. Depth estimation error

Median (cm)	15.6
Average (cm)	18.1
Standard deviation (cm)	13.5
Median (%)	3.2
Average (%)	3.8
Standard deviation (%)	2.6

The Vicon motion capture system provides the ground truth and enables the distance error measurement between the wheelchair and the doors at high speed.

Our door set has been placed along a circular arc (see Figure 4), with a view to assessing the capacity detection when multiple elements are present simultaneously on an image. In this use case, the wheelchair moves around the stage and, in changing direction, is repeatedly found on the different doors. The error between the estimated distance after detection by YOlOv3, and the ground truth provided by the Vicon motion capture system. Figure 5 shows the different quantitative results of the actual distance and the distance estimated by the D435 RealSense camera, as well as the difference between these two values. The comparison takes into account a total of 650 door detections. Table 1 summarizes the numerical results of this experiment. It shows an average error of 18.1 cm representing an error of 3.8% in the object's distance estimation. However, these values remain lower than the data provided by the manufacturer.

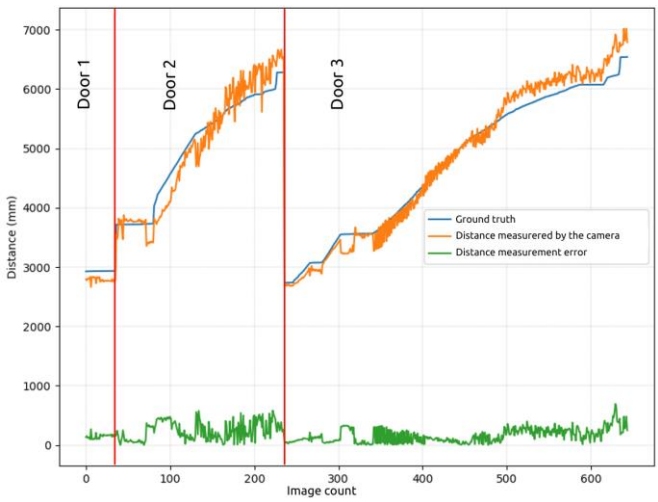

Figure 5. Quantitative results of depth estimation by the RealSense D435 camera. The measured distance between door and camera based on depth images (orange), ground truth (blue), and the difference between both ground truth and camera measurement (green)

6. CONCLUSION

In this paper, we presented objects detection, depth

estimation, localization, and tracking-based deep learning for wheelchair smart mobility. Object detection is based on the YOLOv3 approach. We measured the distance estimation error with the detected objects. Finally, we have improved a version of the SORT algorithm to perform object tracking. Object position estimation is improved by using the extended Kalman filter. To validate the whole of our developments, our models have been re-trained using an open and internal dataset composition: MCIndoor20000 and ESIGELEC datasets. Object detection and tracking were evaluated using the ESIGELEC dataset to validate the wheelchair object detection, localization, and tracking in an indoor environment. Bu having YOLOv3 re-trained on our own dataset, we get good performance in the wheelchair's indoor environment. All developments are integrated on the smart wheelchair platform via an Nvidia Jetson TX2 board playing the role of the main computer. All deep learning algorithms such as object detection, distance estimation, and tracking sun on this same embedded platform.

In future studies, we will develop a new approach-based semantic segmentation to make a good analysis and understanding of the wheelchair outdoor environment. For this, we need to develop a new dataset of the wheelchair outdoor environment of street scenes taken from viewpoints located on sidewalks. It will be the first dataset for wheelchair smart mobility on pathways. We will also develop a new architecture of CNN with temporal processing to improve tracking accuracy.

ACKNOWLEDGMENT

The work presented in this document is supported by the ADAPT project which is carried out within the framework of the INTERREG VA FMA ADAPT project "Assistive Devices for empowering disAbled People through robotic Technologies" http: //adapt-project.com/index.php. The Interreg FCE program is a European Territorial Cooperation Program that aims to fund high-quality cooperation projects in the Channel border region between France and England. The program is funded by the European Regional Development Fund (ERDF)). We would like to thank the engineers of the Autonomous Navigation Laboratory (LNA) of IRSEEM for their support and the Engineering Research and Development Department (SIRD) of ESIGELEC for their help in the test phase. This work benefited from the computing resources of the CRIANN mesocentre (Centre Régional Informatique et d'Applications Numériques de Normandie).

REFERENCES

[1] Girshick, R. (2015). Fast R-CNN. 2015 IEEE International Conference on Computer Vision (ICCV), pp. 1440-1448. http://dx.doi.org/10.1109/ICCV.2015.169

[2] Ren, S., He, K., Girshick, R., Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Trans Pattern Anal Mach Intell., 39(6): 1137-1149. https://doi.org/10.1109/TPAMI.2016.2577031

[3] He, K., Gkioxari, G., Dollár, P., Girshick, R. (2017). Mask R-CNN. IEEE Transactions on Pattern Analysis and Machine Intelligence, 42(2): 386-397. http://dx.doi.org/10.1109/TPAMI.2018.2844175

[4] Pellegrini, T., Fontan L., Sahraoui, H. (2016). Réseau de neurones convolutif pour l'évaluation automatique de la prononciation. Conference: 31ème Journées d'Études sur la Parole, Paris.

[5] Redmon, J., Farhadi, A. (2018). YOLOv3: An incremental improvement. arXiv:1804.02767.

[6] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C., Berg, A.C. (2016). SSD: Single shot multibox detector. In: Leibe B., Matas J., Sebe N., Welling M. (eds) Computer Vision – ECCV 2016. ECCV 2016. Lecture Notes in Computer Science, vol 9905. Springer, Cham. https://doi.org/10.1007/978-3-319-46448-0_2

[7] Hassan, M. (2018). VGG16: Convolutional network for classification and detection.

[8] Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Li, F.F. (2009). Imagenet: A large scale hierarchical image database. 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248-255. http://dx.doi.org/10.1109/CVPR.2009.5206848

[9] Jeong, H.J., Park, K.S., Ha, Y.G. (2018). Image preprocessing for efficient training of yolo deep learning networks. 2018 IEEE International Conference on Big Data and Smart Computing (BigComp), pp. 635-637. http://dx.doi.org/10.1109/BigComp.2018.00113

[10] Yang, J., Zhao, W.L., Deng, C.H., Wang, H.Z., Moon, S. (2017). Fast nearest neighbor search based on approximate k-NN graph. In: Huet B., Nie L., Hong R. (eds) Internet Multimedia Computing and Service. ICIMCS 2017. Communications in Computer and Information Science, vol 819. Springer, Singapore. http://dx.doi.org/10.1007/978-981-10-8530-7_32

[11] Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Kolesnikov, A., Duerig, T., Ferrari, V. (2018). The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. International Journal of Computer Vision, 128: 1956-1981. http://dx.doi.org/10.1007/s11263-020-01316-z

[12] Doyle, S., Feldman, M.D., Shih, N., Tomaszewski, J., Madabhushi, A. (2012). Cascaded discrimination of normal, abnormal, and confounder classes in histopathology: Gleason grading of prostate cancer. BMC Bioinformatics, 13(1): 282. http://dx.doi.org/10.1186/1471-2105-13-282

[13] Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L. (2014). Microsoft COCO: Common objects in context. In: Fleet D., Pajdla T., Schiele B., Tuytelaars T. (eds) Computer Vision – ECCV 2014. ECCV 2014. Lecture Notes in Computer Science, vol 8693. Springer, Cham. http://dx.doi.org/10.1007/978-3-319-10602-1_48

[14] Chen, X., fang, H., Lin, T.Y., Vedantam, R., Gupta, S., Dollár, P., Zitnick, C.L. (2015). Microsoft COCO captions: Data collection and evaluation server. arXiv preprint arXiv:1504.00325.

[15] Caesar, H., Uijlings, J., Ferrari, V. (2018). COCO-stuff: Thing and stuff classes in context. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1209-1218. http://dx.doi.org/10.1109/CVPR.2018.00132

[16] Li, J., Raventos, A., Bhargava, A., Tagawa, T., Gaidon, A. (2018). Learning to fuse things and stuff. arXiv preprint arXiv :1812.01192.

[17] Vicente, S., Carreira, J., Agapito, L., Batista, J. (2014). Reconstructing pascal voc. 2014 IEEE Conference on Computer Vision and Pattern Recognition, pp. 41-48. http://dx.doi.org/10.1109/CVPR.2014.13

[18] Zhu, C., Bichot, C., Chen, L. (2010). Multi-scale color local binary patterns for visual object classes recognition. 2010 20th International Conference on Pattern Recognition, pp. 3065-3068. http://dx.doi.org/10.1109/ICPR.2010.751

[19] Krizhevsky, A., Nair, V., Hinton, G. Cifar-10 and cifar-100 datasets. URl: https://www.cs.toronto.edu/~kriz/cifar.html, accessed on 12 June 2021.

[20] Cui, Y., Jia, M., Lin, T., Song, Y., Belongie, S. (2019). Class-balanced loss based on effective number of samples. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 9260-9269 http://dx.doi.org/10.1109/CVPR.2019.00949

[21] Everingham, M., Van Gool, L., Williams, C. (2007). The PASCAL visual object classes challenge 2007 (VOC2007) results.

[22] Griffin, G., Holub, A., Perona, P. (2007). Caltech-256 object category dataset.

[23] Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Li, F.F. (2019). Imagenet: A large-scale hierarchical image database. 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248-255. http://dx.doi.org/10.1109/CVPR.2009.5206848

[24] Bashiri, F.S., LaRose, E., Peissig, P., Tafti, A.P. (2018). MCIndoor20000: A fully-labeled image dataset to advance indoor objects detection. Data in Brief, 17: 71-75. http://dx.doi.org/10.1016/j.dib.2017.12.047

[25] Bewley, A., Ge, Z., Ott, L., Ramos, F., Upcroft, B. (2016). Simple online and realtime tracking. 2016 IEEE International Conference on Image Processing (ICIP), pp. 3464-3468. http://dx.doi.org/10.1109/ICIP.2016.7533003

[26] Wojke, N., Bewley, A., Paulus, D. (2017). Simple online and realtime tracking with a deep association metric. 2017 IEEE International Conference on Image Processing (ICIP), pp. 3645-3649. http://dx.doi.org/10.1109/ICIP.2017.8296962

[27] Merriaux, P., Dupuis, Y., Boutteau, R., Vasseur, P., Savatier, X. (2017). A study of Vicon system positioning performance. Sensors, 17(7): 1591. http://dx.doi.org/10.3390/s17071591