Long-range obstacle detection from a monocular camera

Muhammad Abdul Haseeb
Institute of Automation
University of Bremen
Bremen, Germany
haseeb@iat.uni-bremen.de

Danijela Ristić-Durrant
Institute of Automation
University of Bremen
Bremen, Germany
ristic@iat.uni-bremen.de

Axel Gräser
Institute of Automation
University of Bremen
Bremen, Germany
ag@iat.uni-bremen.de

ABSTRACT

Reliable and accurate detection of obstacles is one of the challenges of safe autonomous driving. In the past decades, significant work has been done to address the autonomous obstacle detection in different application fields and scenarios [1][2]. In recent years, there is a tendency to use experience from obstacle detection both in the automotive and the aviation sector for the development of autonomous obstacle detection in railways [3][4]. While the main principle of obstacle detection in front of a vehicle from the automotive sector can be applied to railway applications, there are also specific challenges. One of the key challenges is long-range obstacle detection. Sensor technology in current land transport research is able to look 200 m ahead [5]. The required rail obstacle detection interfacing with locomotive control should be able to look ahead up to 1000 m detecting objects on and near track [6]. In this paper, a novel obstacle detection system, which learns and predicts the distance between the object and the camera sensor, is presented. It is based on Multi Hidden-Layer Neural Network, named DisNet, which was trained using a supervised learning technique where the input features were manually calculated parameters of the bounding boxes of objects detected in camera images and outputs were the accurate 3D laser scanner measurement of the distances to objects in the recorded scene. The presented DisNet-based distance estimation system was evaluated on the real-world images of railway scenes acquired by RGB and thermal monocular cameras. Shown results demonstrate DisNet ability for long-range object detection and distance estimation as well as general nature of the proposed system enabling its use for the estimation of distances to objects imaged with different types of monocular cameras.

KEYWORDS

Autonomous obstacle detection on rail tracks. Long-range obstacle distance estimation.

1 DisNet: learning and predicting the distance

DisNet distance estimation system is based on learning the change in object appearance in an image (in terms of size) due to the change of the object distance with respect to camera viewing the object. Fig. 1 illustrates the system architecture. The camera image is the input to the state of the art YOLO object classifier [7] trained with COCO dataset [8]. YOLO is a fast and accurate object detector based on Convolution Neural Network (CNN). Its outputs are bounding boxes of detected objects in the image and labels of the classes of detected objects. The objects bounding boxes resulted from the YOLO object classification are processed to calculate the features-bounding boxes parameters. Based on the input features, the trained DisNet gives as outputs the estimated distance of the object to the camera coordinate system. In Fig. 1, an example of the estimation of distances of two persons on the rail tracks is shown.

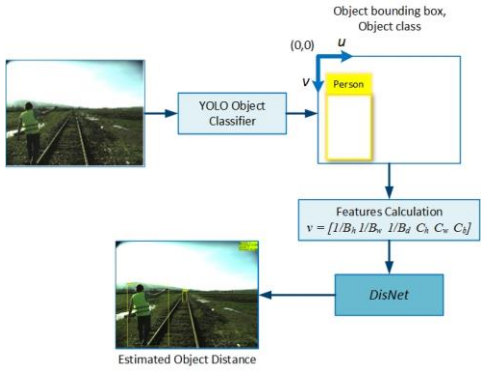


Figure 1: The DisNet -based system used for object distance estimation from a monocular camera

DisNet Training - 2000 features vectors **v** dataset was created by calculation of the parameters of manually extracted objects bounding boxes in RGB images: B_h=(height of the object bounding box in pixels/image height in pixels); B_w=(width of the object bounding box in pixels/image width in pixels); B_d=(diagonal of the object bounding box in pixels/image diagonal in pixels)
Calculated features vectors, **v** have 6 coordinates:

$$\mathbf{v} = [1/B_h \; 1/B_w \; 1/B_d \; C_h \; C_w \; C_b] \quad (1)$$

Besides the inverse of the above bounding boxes parameters, features are C_h, C_w and C_b that represent average height, width and breadth of the particular object class. For example for the class "person" C_h, C_w and C_b are 175 cm, 55 cm and 30 cm respectively. The images used for extraction of features vectors were captured by RGB camera. In order to achieve sufficient discriminatory information in the dataset, different objects, which could be present in a railway scene as possible obstacles on the rail tracks such as pedestrians and bicycles were recorded. The objects position was recorded also with a 3D laser scanner simultaneously, which was placed inline with the camera, on the same distance from the imaged objects and on the same elevation as the camera.

M. A. Haseeb, D. R. Durrant and A. Gräser are with the Institute of Automation, University of Bremen, Otto-Hahn-Allee, NW1, Bremen 28359, Germany (Corresponding author. Tel.:+49-421-218-62446; Fax: +49-421-218-9862446; e-mail: haseeb@iat.uni-bremen.de).

The input dataset was randomly split into a training (80% of the data), validation (10% of the data) and a test set (10% of the data). For the training of DisNet, a feature vector **v** feed to the network input and output, i.e. the ground truth was the accurate object distance measurement from the laser scanner in the recorded scene.

DisNet Structure - The DisNet is a Neural Network with 3 hidden layers with 100 hidden neurons per layer as shown in Fig. 2.

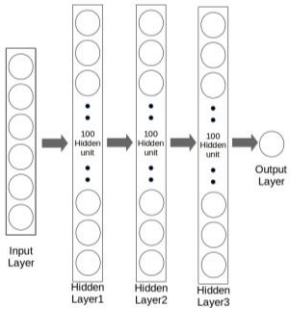


Figure 2: The structure of DisNet - object distance estimation

2 Evaluation

The sensor data, which were used for the evaluation of the proposed *DisNet*-based system for object distance estimation, were recorded in the field tests in different times of the day and night on the location of the straight rail tracks (Fig. 3). During the performed field tests, the members of the author's research team imitated potential obstacles on the rail tracks located on different distances from the sensors test stand.

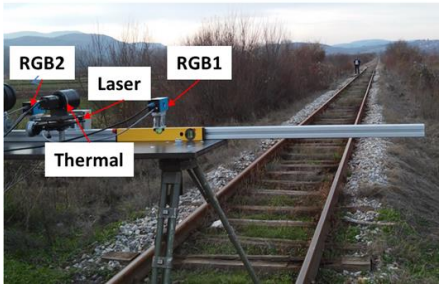


Figure 3: Field tests performed on the straight rail tracks; Test stand with the inline sensors viewing the rail tracks and an object (person) on the rail track

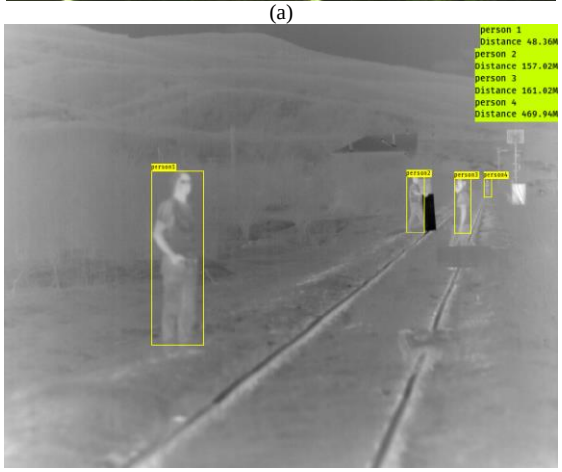
2.1 Distance estimation from single RGB and thermal camera

Some of the results of the DisNet object detection in RGB and thermal camera images are given in Fig. 4. The estimated distances to the detected objects (persons) are given in Table I. Shown result demonstrate long-range distance estimation of up to 500 meters. Moreover, the results show the advantages of multiple viewing angles due to different positioning of cameras on the test-stand. The multiple perspectives assist to detect the person 4 in RGB image, which cannot be seen in the thermal image due to its position behind the person 5. Similarly, person 5 can be seen in the thermal image, but not in the RGB image.

TABLE I. ESTIMATED DISTANCES VS. GROUND TRUTH (IN METERS)

Object	*Ground Truth*	*RGB Camera*	Thermal Camera
Person 1	50	54.26	48.36
Person 2	100	132.26	161.02
Person 3	150	167.59	157.02
Person 4	300	338.51	not-visible
Person 5	500	not-visible	469.94

(a)



(b)

Figure 4: DisNet estimation of distances to objects at different distances in a rail track scene captured by RGB camera (a) and thermal camera (b).

A significant error in estimation of the distance of the person at 100m is due to partially extracted person bounding box by YOLO object classifier. In future work, bounding box extraction will be improved by a combination of YOLO, i.e. machine learning, object detection with traditional image processing based object detection. In addition, other visual features will be investigated to improve the accuracy of distance estimation.

ACKNOWLEDGMENT

This research has received funding from the Shift2Rail Joint Undertaking under the European Union's Horizon 2020 research and innovation programme under grant agreement No 730836.

REFERENCES

[1] N. Bernini, M. Bertozzi, L. Castangia, M. Patander, M. Sabbatelli, Real-Time Obstacle Detection using Stereo Vision for Autonomous Ground Vehicles: A Survey, 2014 IEEE 17th International Conference on Intelligent Transportation Systems (ITSC), China.

[2] A. Geiger, P. Lenz and R. Urtasun, Are we ready for autonomous driving? The KITTI vision benchmark suite, in IEEE Int. Conf. on Computer Vision and Pattern Recognition, 2012.

[3] A. Berg, K. Oefjaell, J. Ahlberg, M. Felsberg, Detecting Rails and Obstacles Using a Train-Mounted Thermal Camera, R.R. Paulsen and K.S. Pedersen (Eds.): SCIA 2015, LNCS 9127, pp. 492–503, 2015. DOI: 10.1007/978-3-319-19665-7 42.

[4] J. Weichselbaum, C. Zinner, O. Gebauer, W. Pree, Accurate 3D-vision-based obstacle detection for an autonomous train, Computers in Industry 64 (2013), pp. 1209–1220.

[5] P. Pinggera, U. Franke, R. Mester, High-performance long range obstacle detection using stereo vision, 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS2015).

[6] Shift2Rail Joint Undertaking, Multi-Annual Action Plan, Brussels, November 2015.

[7] Redmon, Joseph and Farhadi, Ali, *YOLOv3: An Incremental Improvement*, arXiv, 2018.

[8] Lin TY. et al. (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