DisNet: A novel method for distance estimation from monocular camera*

Muhammad Abdul Haseeb, Jianyu Guan, Danijela Ristić-Durrant, Axel Gräser, *Member, IEEE*
Institute of Automation, University of Bremen, Otto-Hahn-Allee NW1, 28359 Bremen, Germany

***Abstract*— In this paper, a machine learning setup that provides the obstacle detection system with a method to estimate the distance from the monocular camera to the object viewed with the camera is presented. In particular, the preliminary results of an on-going research to allow the on-board multisensory system, which is under development within H2020 Shift2Rail project SMART, to autonomously learn distances to objects, possible obstacles on the rail tracks ahead of the locomotive are given. The presented distance estimation system is based on Multi Hidden-Layer Neural Network, named DisNet, which is used to learn and predict the distance between the object and the camera sensor. The DisNet was trained using a supervised learning technique where the input features were manually calculated parameters of the object bounding boxes resulted from the YOLO object classifier and outputs were the accurate 3D laser scanner measurements of the distances to objects in the recorded scene. The presented DisNet-based distance estimation system was evaluated on the images of railway scenes as well as on the images of a road scene. Shown results demonstrate a general nature of the proposed DisNet system that enables its use for the estimation of distances to objects imaged with different types of monocular cameras.**

I. Introduction

Reliable and accurate detection of obstacles is one of the core problems that need to be solved to enable autonomous driving. In the past decades, significant work has been done to address the problem of obstacle detection [1][2]. Besides the emerging of novel algorithms, technology development also enables progress in autonomous obstacle detection. Different onboard sensors such as radars, mono/stereo cameras, LIght Detection And Ranging - LiDAR, ultrasonic sensors and others, implemented in so-called Advanced Driving Assistance Systems (ADAS), are rapidly increasing the vehicle's automation level [3][4].

Many approaches have been presented for different application fields and scenarios. Whereas other transport modes have been quick to automate certain operations, rail runs the risk of lagging behind. One of the key challenges, which has so far hindered automation of rail systems, is the lack of a safe and reliable onboard obstacle detection system for trains within existing infrastructure [5]. 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 [6]. 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 some 200 m ahead [7]. The required rail obstacle detection interfacing with loco control should be able to look ahead up to 1000 m detecting objects on and near track which may potentially interfere with the clearance and ground profile.

The method for long-range obstacle detection presented in this paper is developed within project "SMART-SMart Automation of Rail Transport", funded by the Shift2Rail Joint Undertaking under the European Union's Horizon 2020 research and innovation programme [8]. The main goal of this project is to increase the effectiveness and capacity of rail freight through the contribution to automation of railway cargo haul at European railways by developing of a prototype of an autonomous Obstacle Detection System (ODS). Project SMART will contribute to the long-term vision for an autonomous rail freight system, by the development, implementation and evaluation of a prototype integrated on-board multi-sensor system for reliable autonomous detection of potential obstacles on rail tracks, which could assist drivers and in long term could be used for autonomous initialization of braking of the freight train.

As illustrated in Fig. 1, the SMART ODS combines different vision technologies: thermal camera, night vision sensor (camera augmented with image intensifier), multi stereo-vision system (cameras C1, C2 and C3) and laser scanner (LiDAR) in order to create a sensor fusion system for mid (up to 200 m) and long range (up to 1000 m) obstacle detection, which is independent of light and weather conditions.

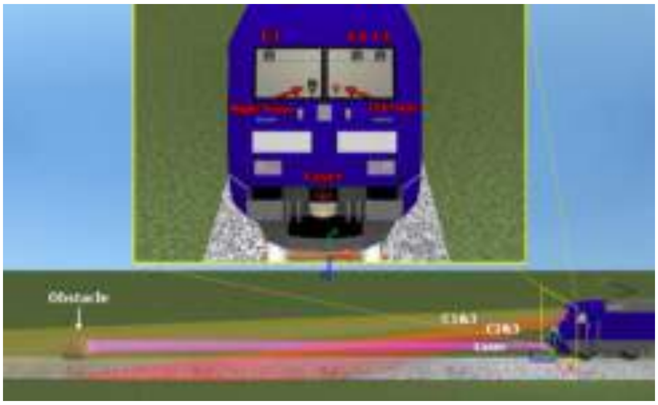


Figure 1. Concept of the SMART multi-sensor ODS. (Top) Front view of the sensors mounted on a locomotive. (Bottom) Side view of the range sensors and an obstacle detection scene.

*Research supported by Shift2Rail Joint Undertaking under the European Union's Horizon 2020 research and innovation programme under grant agreement No 730836.

M. A. Haseeb, J. Guan, D. Ristić-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 main idea behind the multi-sensory system is to fuse the sensor data as sensors individually are not yet powerful enough to deal with complex obstacle detection tasks in all the SMART defined application scenarios, which include day and night operation and operation in poor visibility condition. Because of this, the development of an adequate data fusion system, which effectively combines data streams from multiple sensors, is required. The data fusion approach will be designed based on sensor data availability. Namely, independently of the illumination condition, sensor data from the thermal camera and laser scanner will be always available, where the implemented laser scanner data will be reliably available only in the certain range of up to 100 m. In contrast to that, the stereo camera system fails to generate data under poor illumination conditions, and the night vision camera cannot operate during the day. After obtaining fused data, based on the individual advantages of each sensor, the resulting data stream will be used for detection of obstacles on the rail tracks and for calculation of the distances from the locomotive to detected obstacles. While for stereo cameras traditional depth extraction can be used for thermal camera and night vision camera, estimation of distances from single camera shall be performed.

In this paper, initial results on object distance estimation from monocular cameras are shown using a novel machine learning based method named as DisNet – a multilayer neural network for distance estimation. Although the presented method has been originally developed for autonomous obstacle detection in railway applications, it can be applied to road scenes as well, as it is illustrated in the evaluation section of this paper.

II. Related Work

One of the crucial tasks in autonomous obstacle detection nowadays is finding the solutions to the combination of the environment perception sensors, where vision-based obstacle detection is still considered irreplaceable [9]. Besides its cheaper price, vision is also known as much evolving technology where most of its data is usable as compared to radar and LiDAR [10].

Obstacle detection in computer vision is most commonly done via stereo vision, in which images from two stereo cameras are used to triangulate and estimate distances to objects, potential obstacles, viewed by cameras [11]. Besides the individual use of stereo vision, in a number of obstacle detection systems stereo vision is combined with other range sensors. For example, in [3], an obstacle detection system was developed based on a fusion system consisting of computer vision and laser scanner. The laser provided a point cloud (PC) from which the system extracted the obstacles (clusters of points). These clusters were used both for the region of Interest (ROI) generation for computer vision and as information for obstacle classification, based on machine learning.

Beyond stereo/triangulation cues, there are also numerous monocular cues such as texture variations and gradients, defocus, and colour/haze, which contain useful and important depth information. Some of these cues apply even in regions without texture, where stereo would work poorly. Because of this, some authors follow the idea of human perception of depth by seamlessly combining many of stereo and monocular cues. In [1], a Markov Random Field (MRF) learning algorithm to capture some of these monocular cues is applied, and cues are incorporated into a stereo system. It was shown that by adding monocular cues to stereo (triangulation) ones, significantly more accurate depth estimates than is possible using either monocular or stereo cues alone is obtained. In [13], supervised learning to the problem of estimating depth maps only from a single still image of a variety of unstructured environments, both indoor and outdoor, was applied. However, depth estimation from a single still image is a difficult task, since depth typically remains ambiguous given only local image features. Thus, the presented algorithm must take into account the global structure of the image, as well as use prior knowledge about the scene.

In this paper, a novel method for object distance estimation from a single image, which does not require either a prior knowledge about the scene or explicit knowledge of the camera parameters, is presented. The presented distance estimation system is based on Multi Hidden-Layer Neural Network, named DisNet, which is used to learn and predict the distance between the object and the camera sensor.

III. Neural Network-Based Object Distance Estimation From Monocular Camera

The architecture of the DisNet-based distance estimation system is illustrated in Fig. 2. The camera image is input to the Object Classifier which is based on a state-of-the-art computer vision object detector YOLO (You Only Look Once) [14] trained with COCO dataset [15]. 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 detected objects belong to. The objects bounding boxes resulted from the YOLO object classification are then 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 sensor. In the system architecture illustrated in Fig. 2, an example of the estimation of distances of two persons on the rail tracks is shown.

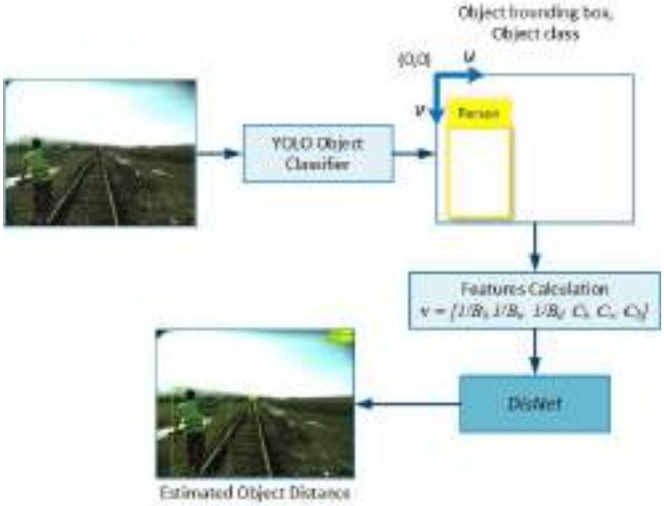


Figure 2. The DisNet -based system used for object distance estimation from a monocular camera

For the training of DisNet, a supervised learning technique was used. This method required a collected dataset including both inputs and outputs, i.e. the ground truth. In the presented system, training dataset was collected manually by manual extraction of 2000 bounding boxes of different objects in the images recorded by RGB cameras at different distances together with the ground truth, which was the accurate laser scanner measurement of the distances to objects in the recorded scene. The details of the structure and training of DisNet are given in the following sections.

A. DisNet training - Dataset

In the presented work, the objective is that DisNet is trained for the estimation of an object's distance to the onboard sensory obstacle detection system. More formally, the task is to estimate the distance to an object in the laser's reference frame, which is on the same distance from the object as the camera reference frame, given an input also called feature vector $\boldsymbol{v}$. In the presented work, $\boldsymbol{v}$ contains the features of the bounding box of the object detected in camera images and the ground-truth is the distance to the object as measured by the laser scanner.

In order to build the dataset, the objects positions and their bounding boxes in the RGB images were manually extracted and 2000 input feature vectors were created. In order to achieve sufficient discriminatory information in the dataset, different objects at different distances, which could be present in a railway scene as possible obstacles on the rail tracks, were considered. Some of the objects recorded at different distances and their bounding boxes from the dataset are shown in Fig. 3.

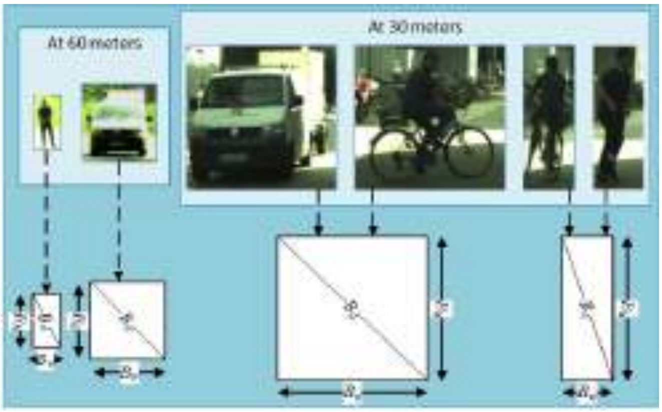


Figure 3. Examples from the DisNet dataset of different object bounding boxes in the RGB images

For each extracted object bounding box, a six-dimensional feature vector $\boldsymbol{v}$ was calculated:

$$\boldsymbol{v} = [1/B_h \ 1/B_w \ 1/B_d \ C_h \ C_w \ C_b] \qquad (1)$$

where the coordinates of vector $\boldsymbol{v}$, features, are:

Height, B_h=(height of the object bounding box in pixels/image height in pixels)
Width, B_w=(width of the object bounding box in pixels/image width in pixels)
Diagonal, B_d=(diagonal of the object bounding box in pixels/image diagonal in pixels)

The ratios of the object bounding box dimensions to the image dimensions B_h, B_w and B_d enable the reusability of DisNet trained model with a variety of cameras independent of image resolution. C_h, C_w and C_b in (1) are the values of average height, width and breadth of an object of the particular class. For example for the class "person" C_h, C_w and C_b are respectively 175 cm, 55 cm and 30 cm, and for the class "car" 160 cm, 180 cm and 400. The features C_h, C_w and C_b are assigned to objects labelled by YOLO classifier as belonging to the particular class in order to complement 2D information on object bounding boxes and so to give more information to distinguish different objects.

The relationships of the calculated features of object bounding boxes in 2D image, B_h, B_w and B_d, and the real distance to the image measured by laser scanner in the range 0-60 m, are given in Fig. 4. Geometrically, by the projective transformations, the object bounding box size is expected to get smaller the further away the object is, so the inverse of bounding box size is expected to increase as the distance increases. Inspection of the data confirms that this is the case and suggests that the relationship is approximately linear, which gives a clear motive to use it for the dataset used for training of DisNet.

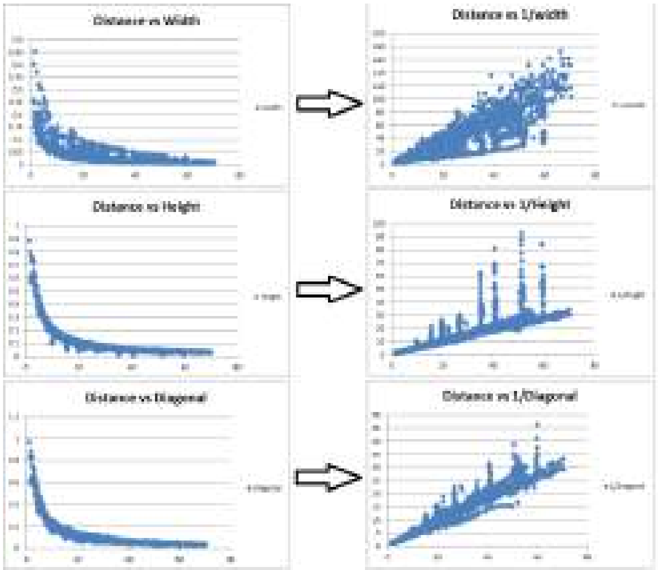


Figure 4. DisNet features vs. distance

For training the network the input dataset was firstly randomly split into a training (80% of the data), validation (10% of the data) and test set (10% of the data).

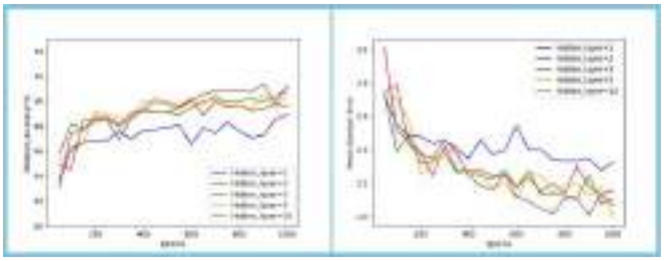
The DisNet was trained using the backpropagation method with Adam optimizer [16] on the dataset collected.

B. DisNet structure

In order to find the appropriate number of hidden layers experiments with various numbers of hidden layers (1,2,3,5 and 10) were performed assuming that each hidden layer had 100 neurons. Fig. 5 (a) shows the accuracy of distance estimation over 1000 epochs achieved for different number of hidden layers. As obvious, DisNet with one hidden layer achieves the lowest distance estimation accuracy. It is also obvious that there is no significant difference in distance estimation accuracy achieved with 2,3,5 and 10 hidden layers. For this analysis, a reduced dataset was used. The networks were trained on the 80% dataset and the estimation accuracy reported is on the 10% validation set.

Similar behaviour can also be seen in Fig. 5(b) where the Mean Absolute Error over 1000 epochs achieved for a different number of hidden layers is shown. As obvious, the Mean Absolute Error is largest for the DisNet with one hidden layer, while there is no significant difference in the Error achieved with 2,3,5 and 10 hidden layers.

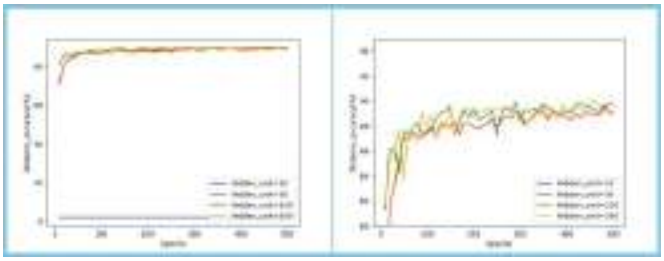
Even though the smallest values of Mean Absolute Error were achieved for 10 hidden layers and the distance accuracy was highest for 10 hidden layers, a trade-off was made between the computational time and accuracy/error and finally, DisNet with 3 hidden layers was chosen.



(a) (b)

Figure 5. (a) Distance Estimation Accuracy and (b) Mean Absolute Error achieved for different numbers of hidden layers

After making a decision on network with 3 hidden layers, in order to find the appropriate number of neurons for the hidden layers experiments with various numbers of hidden neurons were performed. Fig. 6 (a) shows the accuracy of distance estimation over 1000 epochs achieved for different number of neurons per hidden layer. As obvious, the distance estimation accuracy achieved with 10 hidden neurons is very low, much lower than distance estimation accuracy achieved with 30, 100 and 200 hidden neurons. The magnified diagram in Fig. 6 (b) shows that distance estimation accuracy with 30 hidden neurons is lower than with 100 and 200 neurons. Bearing in mind that there is no significant difference in distance accuracy estimation with 100 and 200 hidden neurons, in order to reduce the complexity of DisNet, finally, 100 neurons per hidden layer were chosen.



(a) (b)

Figure 6. Distance Estimation Accuracy achieved for different number of hidden neurons per hidden layer in 3-hidden layers neural network DisNet

The final structure of DisNet having 3 hidden layers with 100 hidden neurons per layer is shown in Fig. 7.

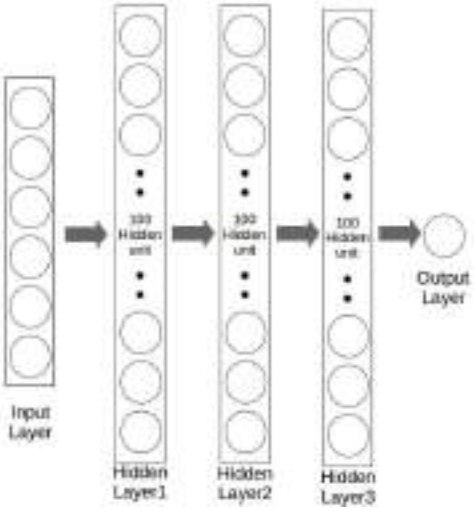


Figure 7. The structure of DisNet used for object distance prediction

DisNet input layer consists of 6 neurons corresponding to 6 features, parameters of output layer consists of only one single neuron. The output of this node is the estimated distance between the camera and the object viewed with the camera.

IV. EVALUATION

The DisNet-based system for distance estimation was evaluated on images recorded in the field tests within the H2020 Shift2Rail project SMART [8]. The sensor data, which were used for the evaluation of a DisNet-based system for object distance estimation, were recorded in the field tests on the straight rail tracks in different times of the day and night on the location of the straight rail tracks (Fig. 8). Monocular RGB cameras were mounted on the static test-stand, together with the laser scanner in the locations which resemble their intended locations in the final integrated SMART obstacle detection (Fig. 8). During the performed field tests, the members of the SMART Consortium imitated potential static obstacles on the rail tracks located on different distances from the SMART test-stand.

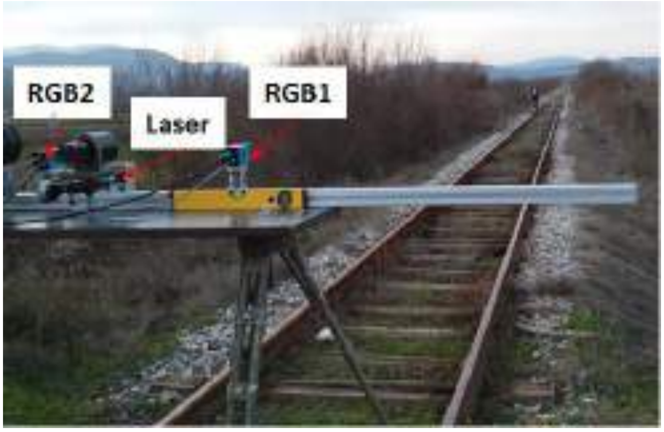


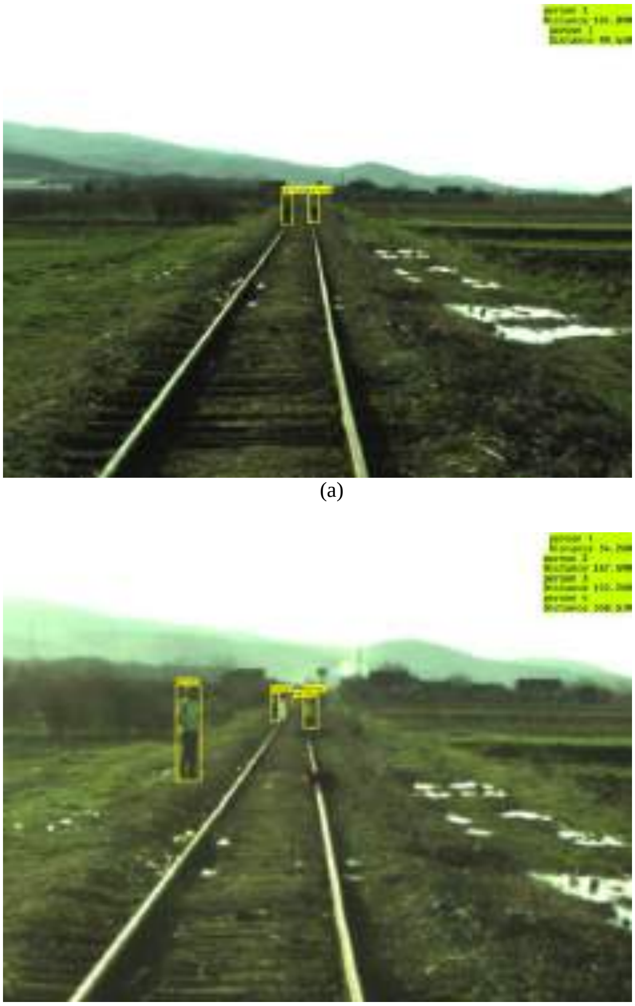
Figure 8. Field tests performed on the straight rail tracks; Test-stand with the SMART sensors viewing the rail tracks and an object (person) on the rail track

A. Railway Scene - Distance estimation from the single RGB camera image

Some of the results of the DisNet object distance estimation in RGB images are given in Fig. 9. The estimated distances to the objects (persons) detected in the images are given in Table I.

TABLE I. ESTIMATED DISTANCES VS. GROUND TRUTH

Figure	**Object**	**Rail Scene**	
		Ground Truth	***Distance estimated from DisNet***
9 (a)	Person 1	100 m	101.89 m
	Person 2		99.44 m
9 (b)	Person 1	50 m	54.26 m
	Person 2	150 m	167.59 m
	Person 3	100 m	132.26 m
	Person 4	300 m	338.51 m



(a)

(b)

Figure 9. DisNet estimation of distances to objects in a rail track scene from the RGB camera image. (a) Distance estimation of detected persons at 100 m and (b) Magnified RGB image overlaid with bounding boxes and distance estimation of detected persons at 50, 100, 150 and 300 m respectively.

As obvious from Fig. 9, YOLO based object detection in images is reliable in spite of the fact that YOLO classifier was used in its original form trained with COCO dataset [15], without re-training with the images from the SMART field tests. Also, it is obvious that achieved distance estimation is satisfactory in spite of the fact that DisNet database did not contain object boxes from the real rail tracks scenes. This, in the first place, means that the objects in real field tests scenes were at larger distances from the sensors than in the recording tests used for dataset building. Also, the distances of the objects in field tests were outside the laser scanner range used for the training of DisNet. The difference in estimation of distances of persons at 100 m (Fig. 9(a) and 9(b)) indicates the need for improvement of objects bounding boxes extraction. Namely, the person at 100 m in Fig. 9(b) is not fully bounded with the bounding box as the lower part of the person is occluded by a board. Also, in future work, novel features with a higher correlation score with respect to distance will be investigated and will be used to improve the accuracy of distance estimation in SMART obstacle detection system.

A. Road Scene - Distance estimation from the single RGB camera image

Although presented DisNet-based method for distance estimation from the monocular camera has been originally developed for autonomous obstacle detection in railway applications, it can be applied to road scenes as well. To demonstrate this, presented method was applied to the image of a different resolution than images used for training of DisNet. The image of a road scene was recorded within the project HiSpe3D-Vision presented in [11][17]. The main goal of HiSpe3D-Vision was to develop a high speed, low latency stereo vision based collision warning system for automotive applications. The obstacle detection and distance calculation for collision warning were based on the segmentation of the disparity map created from the car-mounted stereo-vision system. The result of object detection and distance estimation in a dynamic environment (moving car and moving object-obstacle) is shown in Fig. 10, where original image is overlaid with the bounding cuboid for the object closest to the car (person on the bike). Distance for this object, as estimated by the HiSpe3D-Vision method, is given in the left upper corner of the image in Fig. 10, as well as in Table II.

In contrast to HiSpe3D-Vision method, which detected only the object closest to the car, the presented DisNet method recognized different objects in the scene recorded by the car-mounted camera: person, bicycle, car and track. The bounding boxes of the recognized objects are overlaid on the image in Fig. 10 together with distances estimated by DisNet. The objects distance estimation achieved by DisNet vs. the distance estimation achieved by HiSpe3D stereo vision method is given in Table II.

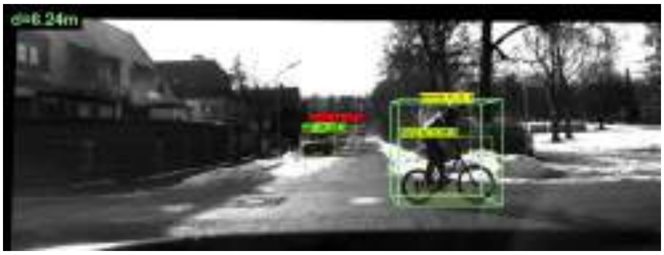


Figure 10. Road scene image overlaid with objects recognition and distance estimation results achieved by proposed DisNet and by stereo-vision based HiSpe3D method [17]

TABLE II. OBJECTS DISTANCES ESTIMATED BY DISNET VS. OBJECTS DISTANCES ESTIMATED BY HISPE3D-VISION METHOD [17]

Object	**Road Scene**	
	Distance estimated by HiSpe3D-Vision	***Distance estimated by DisNet***
Person	6.24 m	6.12 m
Bicycle	-	5.39 m
Car	-	27.64 m
Truck	-	30.25 m

As obvious, DisNet outperforms HiSpe3D-Vision method in a number of different objects recognized in the recorded scene. The person distance estimation by both methods is comparable. The difference in distances for the person and the bicycle, estimated by DisNet, indicates the need for improvement of objects bounding boxes extraction. In future work, the YOLO classifier will be replaced with 2D image processing based classifier and bounding box extractor, which is under development in the SMART project.

V. CONCLUSION

In this paper, the initial results of DisNet – a machine learning-based distance estimation from the monocular camera, achieved by obstacle detection system under development within Shift2Rail project SMART-Smart Automation of Rail Transport, are presented. Presented results illustrate reliable estimation of distances from a single RGB camera to objects in static railway scenes recorded by cameras. General nature of the presented distance estimation method is demonstrated by the result of distance estimation in a dynamic road scene captured with different types of cameras. This indicates that in future work presented method can be used for object distance estimation from different types of monocular cameras integrated into the SMART on-board obstacle detection system, thermal camera and night vision camera. Further, the presented obstacle detection system will be evaluated in dynamic field tests when mounted on a locomotive in motion. The presented results from dynamic road scene justify the expectation that DisNet-based obstacle detection system will work on real experimental images when the train is in motion.

ACKNOWLEDGEMENT

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] F. Jiménez, J. E. Naranjo, J. J. Anaya, F. García, A. Ponz, J. M. Armingol, Advanced Driver Assistance System for road environments to improve safety and efficiency, Transportation Research Procedia 14 (2016) 2245 – 2254.

[4] S. Kim, H. Kim, W. Yoo, K. Huh, Sensor Sensor Fusion Algorithm Design in Detecting Vehicles Using Laser Scanner and Stereo Vision, IEEE Transactions on Intelligent Transportation Systems, Vol. 17, No. 4, 2016.

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

[6] 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.

[7] 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).

[8] Shift2Rail project SMART: http://www.smartrail-automation-project.net

[9] F. de Ponte Müller, Survey on Ranging Sensors and Cooperative Techniques for Relative Positioning of Vehicles, Sensors 2017, 17(2).

[10] J. Park, J.-H. Lee, S. H. Son, A Survey of Obstacle Detection using Vision Sensor for Autonomous Vehicles, 2016 IEEE 22nd International Conference on Embedded and Real-Time Computing Systems and Applications.

[11] A. Leu, D. Aiteanu, A. Gräser, High Speed Stereo Vision Based Automotive Collision Warning System, Applied Computational Intelligence in Engineering and Information Technology, Volume 1, pp. 187-199, Springer Verlag Heidelberg, 2012.

[12] A. Saxena., J. Schulte, A. Y. Ng, Depth estimation using monocular and stereo cues. In: IJCAI, 2007.

[13] A. Saxena, H.·Sung, A. Y. Ng, 3-D Depth Reconstruction from a Single Still Image; Int J Comput Vis, 2007.

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

[15] COCO dataset, https://arxiv.org/pdf/1405.0312.pdf

[16] D. P. Kingma, J. L.Ba, ADAM: A Method for Stochastic Optimization,https://arxiv.org/pdf/1412.6980.pdf

[17] A. Leu, D. Bacără, D. Aiteanu, A. Gräser, Hardware acceleration of image processing algorithms for collision warning in automotive applications, Methods and Applications in Automation: 32nd – 33rd Colloquium of Automation, Salzhausen/Leer, Germany, pp.1-12, Shaker Verlag GmbH, 2012.