CNN-LIDAR pedestrian classification: combining range and reflectance data

Gledson Melotti, Alireza Asvadi, Cristiano Premebida

***Abstract*—The use of multiple sensors in perception systems is becoming a consensus in the automotive and robotics industries. Camera is the most popular technology, however, radar and LIDAR are increasingly being adopted more often in protection and safety systems for object/obstacle detection. In this paper, we particularly explore the LIDAR sensor as an inter-modality technology which provides two types of data, range (distance) and reflectance (intensity return), and study the influence of high-resolution distance/depth (DM) and reflectance maps (RM) on pedestrian classification using a deep Convolutional Neural Network (CNN). Pedestrian protection is critical for advanced driver assistance system (ADAS) and autonomous driving, and it has regained particular attention recently for known reasons. In this work, CNN-LIDAR based pedestrian classification is studied in three distinct cases: (i) having a single modality as input in the CNN, (ii) by combining distance and reflectance measurements at the CNN input-level (early fusion), and (iii) combining outputs scores from two single-modal CNNs (late fusion). Distance and intensity (reflectance) raw data from LIDAR are transformed to high-resolution (dense) maps which allow a direct implementation on CNNs both as single or multi-channel inputs (early fusion approach). In terms of late-fusion, the outputs from individual CNNs are combined by means of non-learning rules, such as: minimum, maximum, average, product. Pedestrian classification is evaluated on a 'binary classification' dataset created from the KITTI Vision Benchmark Suite, and results are shown for the three cases.**

***Index Terms*—Pedestrian classification; Deep learning; LIDAR perception system; active protection systems**

I. Introduction

One of the key components comprised in Autonomous Driving Systems (ADS) is sensory/artificial perception, which in turns encompasses computer vision (here, including LIDAR "vision"), sensor-fusion, and environment representation. Regardless of the sensors and the representation models, the common denominator in a perception system is AI/ML based algorithms, where deep learning has recently gained considerable attention and interest from automotive and robotics industries and academia. This paper approaches LIDAR-based perception for pedestrian classification using Convolutional Neural Networks (CNN), and TensorFlow, as the supervised classifier. Pedestrian detection - which depends on a pedestrian classification technique - is an important topic in the ITS/IV communities and, recently, it regained more attention for obvious reasons [1].

Although the tremendous efforts on pedestrian safety systems and ADAS technology, the number of accidents involving pedestrians is sadly very high. Therefore, the development of more reliable and effective pedestrian detection systems (PDS) is still a key step forward to reduce road and urban accidents. The advances in PDS are remarkable [2]–[5], but it is a long way to obtain a robust solution for all Operational Design Domain (ODD) conditions. Motivated by deep-learning performance and by the importance of perception systems for ADAS and ADS, this paper aims to study LIDAR-based pedestrian classification using CNN, exploring early and late information fusion approaches. The LIDAR is a remote sensing mechanism composed mainly of a laser and scanner (scanning system), which provides $3D$ point-clouds with cartesian coordinates (x, y and z) and also reflectance value (intensity), as shown in Figure 1.

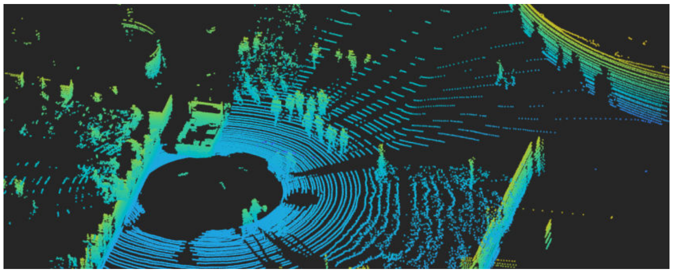

Fig. 1. Example of an image generated by a 3D LiDAR (HDL-64E Velodyne). LIDAR scan obtained from the Object Detection Evaluation of the KITTI dataset [6].

In this work, LIDAR is explored as a multimodal sensor in the sense that the distance (range) and also reflectance (intensity) are both used in the form of high-resolution maps which have the benefit of compensating the low-resolution of a LIDAR [7]. Hereafter, we will refer to these maps as distance/depth map (DM) and reflectance map (RM) as shown in Fig. 2.

Camera based pedestrian classification have been widely addressed by the scientific community [8]–[11]. Active sensors like automotive radar and LIDAR [6], [12], [13], on the other hand, are more robust than cameras w.r.t. illumination changes and also have the pro of measuring distance (a physical property) directly. The drawbacks of LIDAR technologies are that they are expensive, when compared to cameras, and still[1] have moving parts.

Using the range data provided by a LIDAR sensor to obtain depth maps was reported in [7], where the authors

The authors are with the Department of Electrical and Computer Engineering, Institute of Systems and Robotics, University of Coimbra, Portugal. {gledson.melotti,asvadi,cpremebida}@isr.uc.pt.

[1] Some recently launched solid-states LIDAR do not use moving mechanisms.

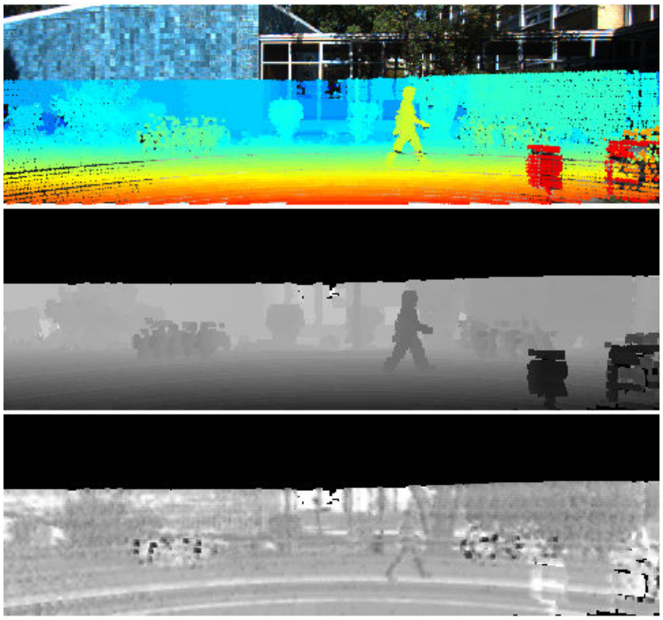

Fig. 2. This picture shows a LIDAR point-cloud (the 1st row) as projected to the image plane (1st row), and the correspondent depth (DM) and reflectance (RM) maps; 2nd and 3rd rows respectively.

benchmarked several techniques of estimating distance (depth) maps by upsampling the LIDAR scans through spatial-filtering by using increasing mask (filter) size. In the possibility of contributing with pedestrian classification using convolutional neural networks, this paper presents a classification study using CNN applied to depth and also reflectance maps generated from a 3*D* LIDAR. In this work we will use a "classification" dataset built from the KITTI suite [6], however camera images will be not used in the CNN classifier. KITTI is a state-of-the-art benchmark for pedestrian detection in urban and road environments.

The possibility of including depth and reflectance maps in the CNN-classification step is studied and we show that the sampling techniques employed to obtain high-resolution maps from LIDAR point-clouds and various mask sizes have different results. In this way the contribution of this paper is a study of CNNs capacity in depth and reflectance maps for pedestrian classification, also the evaluation of the classification using early and late fusion strategies, with learning (CNN) and deterministic techniques (average, minimum, maximum and product).

The structure of this paper is as follows: in Section II related works are revisited. The LIDAR based depth and reflectance maps are explained in Section III. Section IV describes the dataset, while the CNN classification technique is presented in Section V, and sensor fusion approaches are described in Section VI. Results and conclusions are provided in Sections VII and VIII respectively.

II. Related Work

Ensuring the safety of pedestrians on both roads and cities is the focus of many automotive industries and also the scientific community. One way to do this is to develop sensory perception systems capable of classifying/detecting pedestrians, since they are the baseline for ADS and ADAS. Munder and Gavrila [2] explored the combination of SVM (support vector machine), NN (neural networks) and KNN (K-nearest neighbours) and compare global, local, adaptive and non-adaptive analysis (PCA-main component analysis, HW-Haar Wavelets, LRF-local receptive fields). The authors concluded that the performance of global resources is smaller than the local ones and that the adaptive resources are better than the non-adaptive ones. In another study presented by [8], pedestrian classification was performed using data from multiple domains (light spectra: visible and infrared) and multiple modalities (intensity, depth, and motion). In addition to providing a public data set with infrared and visible spectra images, a study on feature extraction of both spectra was obtained by means of LBP (local binary patterns), LGP (local gradient patterns), ISS (intensity self-similarity) and HOG (histogram of oriented gradients), with the purpose of performing pedestrian classification using a SVM linear classifier. The best result was obtained by fusion of the multiple domains (visible and infrared) with the depth modality for each feature extractor, *e.g.*, the fusion was not performed between feature extractors, only between modalities and domains for each extractor.

A recent study in pedestrian detection, this time using CNN, that includes information fusion was presented by [14]. The fusion considers the joining of four independent components (layers): feature extraction, deformation manipulation models, occlusion manipulation models, and classifier. These components interact through a deep model proposed by the authors, in which the layer of deformation is incorporated in a CNN. Through the interaction between these independent components, the result achieved an improvement in accuracy. All these contributions ([2], [8], [14]), however, did not present detection/classification studies using data from by LIDARs sensors, *i.e.*, they applied pedestrian detection and/or classification based on camera images.

Among several techniques to perform the tasks of extracting features and classifying/detecting objects in images, such as pedestrians, the CNN has been demonstrated to be the most efficient in terms of classification performance. The first highly satisfactory result using the CNN with gradient descent was the architecture of the LeNet network presented by [15], which classified manuscript characters with minimum pre-processing. Without including data entry, the network consists of seven layers: three convolution layers, two subsamplings (pooling), a fully connected layer, and the output (last layer) to classify ten classes. CNN technique came to be widely used after the ImageNet ILSVRC challenge in 2012, when AlexNet CNN [16] was ranked in the first place in the challenge. AlexNet is similar to LeNet, but with five layers of convolution: three pooling and two fully connected layers and with an output to classify thousand classes. In addition to these networks, others have been presented by several surveys, from which we cite the best known: ResNet [17], GoogLeNet [18], VGG [19] and ZefNet [20].

Most of the public available datasets on pedestrian classification/detection, see [8] for a review, are image based. KITTI dataset [6], which is the state-of-the-art dataset for urban/road perception, has the advantage of providing synchronized and calibrated data from monocular cameras, stereo-system, and also 3*D* LIDAR scans. Furthermore, it provides examples of “partly occluded”, “fully occlude”, “unknown” and “don’t care” region objects, which make the classification problem more challenging and realistic. KITTI contains several labeled objects, such as pedestrian, car, train and cyclist, for example. The objects are cropped off from annotated images. In this paper we separate into two classes: pedestrian and non-pedestrian, as described in Section IV.

In summary, this paper presents a CNN-LIDAR based pedestrian classification study by means of the depth and reflectance maps generated by upsampling techniques as detailed in Section III. This work differs from the previously cited papers because our CNN classifier uses only data from a LIDAR therefore, RGB data is not feed into the CNN classifiers neither in the fusion strategies.

III. Depth and reflectance maps from LIDAR data

The ability to represent and model scenarios, known as environment representation, is very important for the classification and detection of objects in the environment as perceived by sensors. In this Section, we describe the DM and RM maps representations and the spatial filtering techniques we used to obtain the LIDAR-based maps.

From a 3*D* point-cloud delivered by the LIDAR we can can obtain a 2*D* map in pixel coordinates *i.e.*, assuming a calibrated LIDAR and camera’s setup a transformation from $\mathbb{R}^3$ to the image-plane $\mathbb{R}^2$, where each point is represented by the position in pixel coordinates; However, because of the sparse nature of the LIDAR scans, several pixel positions in the converted map will be unsampled. Therefore, we should estimate the value of ra_i (DM) and/or re_i (RM) in the unsampled locations to obtain a high-resolution representation. An alternative to estimate the unsampled positions is through spatial filtering implemented by a sliding-window filter (mask) technique.

Basically, spatial filters combine the intensity of the group of pixels belonging to a mask $\mathbf{M}$ with a size $n \times n$. Among the numerous possible values for n, this work studied the mask sizes 9×9, 11×11, 13×13 and 15×15. In terms of interpolation/estimation methods, we make use of the average (Ave), minimum (Min) and maximum (Max) operator, as well as the inverse distance weighting (IDW) and the bilateral filter (BF).

Let $\mathbf{x}_0 = (\mathbf{x}, \mathbf{y})_0$ denotes the location of interest, which is the center of $\mathbf{M}$, and r_0^* be the variable to be estimated, *i.e.*, the range (ra_i) or reflectance (re_i) at x_0. Thus, the IDW and BF can be expressed by:

- IDW:

$$r_0^* = \sum_{i=1}^{n} W_i(\mathbf{x}) r_i \qquad (1)$$

where $W_i(\mathbf{x}) = d_i^{-p}$, $d = ||\mathbf{x}_0 - \mathbf{x}_i||$ is a given distance function and p is a power parameter (positive real number).

TABLE I
Summary of the classification dataset

Training set	n# positives = 2827 n# negatives = 29849
Validation set	n# positives = 314 n# negatives = 3316
Testing set	n# positives = 1346 n# negatives = 14213

- BF:

$$r_0^* = \frac{1}{W} \sum_{x_i \in M} G_{\sigma_s}(||\mathbf{x}_0 - \mathbf{x}_i||) \sigma_R(|r_0 - r_i|) \times r_i \qquad (2)$$

where W is a normalization factor that ensures weights sum to one, G_{σ_s} is inversely proportional to the Euclidean distance between the center of the M and the sampled locations x_i, and G_{σ_R} controls the influence of the sampled points based on their values r_i, depending on the case the variable r_i takes the range (for DM) or the reflectance (RM) values.

IV. Classification Dataset

To evaluate the techniques and approaches discussed here, a pedestrian classification dataset was created based on the 2D object-detection dataset of KITTI[2]. The classes are given in the form of 2*D* bounding boxes labeled manually: ‘Pedestrian’, ‘Car’, ‘Truck’, ‘Tram’, ‘Van’, ‘Person (sitting)’, ‘Cyclist’, and ‘Misc’. In this paper the classes were separated in two categories of interest: pedestrian and non-pedestrian *i.e.*, a binary problem. The number of positives examples is 4487 cropped-images (labeled bounding boxes of type ‘Pedestrian’), while the negative class has 47378 cropped-images (types: ‘Cyclist’, ‘Car’, ‘Person (sitting)’, and so on). It was considered 70% for the training set (10% of that for validation) and the remaining 30% for the testing set. Table I gives a summary of the dataset used in this study.

V. Classification using CNN

Among several convolutional neural networks, this paper opted to use AlexNet CNN architecture [16] with some modifications. We used batch normalization in the first two layers, instead of the local normalization scheme, and in the last layer we use the softmax activation function with two classes, instead of 1000 classes, and dropout of 50%. The network was trained from scratch for the pedestrian and non-pedestrian classes. Through the bounding boxes provided by the KITTI dataset, we cropped the objects contained in the depth and reflectance maps images. The objects have different sizes and therefore they have been resized to the same input size of the AlexNet CNN (227×227). The network architecture is shown in Figure 3, where *Ch* is the number of channels, *KS* is the

[2] www.cvlibs.net/datasets/kitti/eval_object.php?obj_benchmark=2d

kernel size (filter), S is the stride, Op is the convolution output, AF is the activation function, N is the normalization function, $Dense$ is the number of neurons in each fully connected layer and γ is the score.

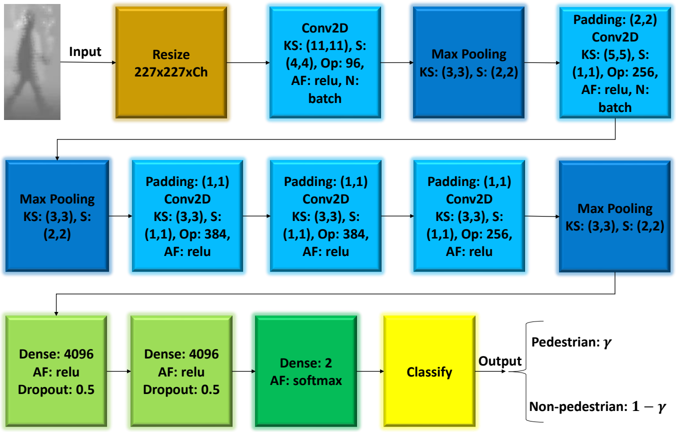
Input
Resize 227x227xCh
Conv2D KS: (11,11), S: (4,4), Op: 96, AF: relu, N: batch
Max Pooling KS: (3,3), S: (2,2)
Padding: (2,2) Conv2D KS: (5,5), S: (1,1), Op: 256, AF: relu, N: batch
Max Pooling KS: (3,3), S: (2,2)
Padding: (1,1) Conv2D KS: (3,3), S: (1,1), Op: 384, AF: relu
Padding: (1,1) Conv2D KS: (3,3), S: (1,1), Op: 384, AF: relu
Padding: (1,1) Conv2D KS: (3,3), S: (1,1), Op: 256, AF: relu
Max Pooling KS: (3,3), S: (2,2)
Dense: 4096 AF: relu Dropout: 0.5
Dense: 4096 AF: relu Dropout: 0.5
Dense: 2 AF: softmax
Classify
Output
Pedestrian: γ
Non-pedestrian: $1-\gamma$

Fig. 3. Modified AlexNet CNN architecture showing the layers of convolution, pooling and classification. Ch is the number of channels, KS is the kernel size (filter), S is the stride, Op is the convolution output, AF is the activation function, N is the normalization function, $Dense$ is the number of neurons in each fully connected layer and γ is the score.

The network was trained with the following parameter settings: 30 epochs, batch size equal 64, stochastic gradient descent optimizer with $lr = 0.001$ (learning rate), $decay = 10^{-6}$ (learning rate decay over each update), $momentum = 0.9$, and categorical cross entropy as loss function.

VI. Sensor Fusion Strategies

The combination, or fusion, of data from distinct sources, in the scope of object recognition, is usually performed by early fusion or late fusion schemes [21]. Depending on the terminology and context under consideration, we can designate such schemes as centralized and decentralized fusion schemes respectively. In the sequel, we describe the way early and later strategies were used to combine data from DM and RM maps for pedestrian classification.

A. Early fusion using CNN

For this case, we trained a CNN with 2 channels where each channel received data from one of the LIDAR modalities: the first channel, in the input layer, of the CNN received distance/depth (DM) and the second was fed with reflectance (RM) maps, as shown in the Figure 4.

B. Late fusion techniques

The number of methods able to combine classifiers outputs is extensive, ranging from non-learning (deterministic) to learning methods. In this work, we will make use of non-learning rules to combine the CNNs outputs, namely: average, maximum, minimum, and product. Here, the latter is used in the form of a normalized-product (which can be understood as a Naive Bayes rule) where additive smoothing is used to prevent non-informative probabilities *i.e.*, to impede values close to zero. The combination of CNN outputs is illustrated in Figure 5.

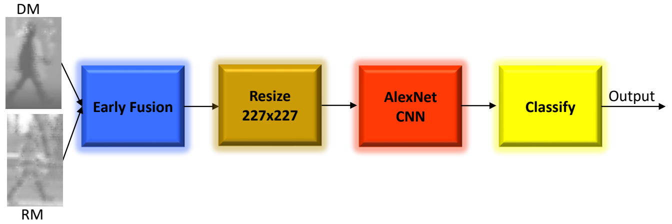

Fig. 4. Early fusion scheme using a 2 channels CNN: a single CNN is trained using one channel for DM and another channel for RM maps.

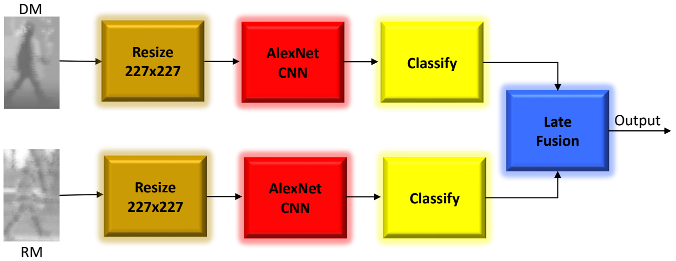

Fig. 5. Late fusion: a CNN is trained with 1 channel for distance/depth (DM) and a second single channel CNN uses reflectance (RM) map. The fusion is obtained later of the classification of each CNN.

Denoting ρ_i the confidence (or probability) score yielded by deep-models CNN_i, $(i = 1, \cdots, m)$, where m is the number of models, CNN_1 denotes a CNN model using DM, and CNN_2 refers to RM (reflectance) CNN-model. Four fusion rules are considered: average $\mathscr{F}_{Mean}$, maximum $\mathscr{F}_{Max}$, minimum $\mathscr{F}_{Min}$, and smooth-product $\mathscr{F}_{Prod}$. The average rule simply calculates the simple mean of the CNN-classifiers outputs $\mathscr{F}_{Mean} = \frac{1}{m}\sum_{i=1}^{m}\rho_i$. The maximum rule outputs the maximum value over the classifier responses, $\mathscr{F}_{Max} = \max\{\rho_i\}$, while the minimum rule is $\mathscr{F}_{Min} = \min\{\rho_i\}$.

Assuming classifiers' independence given the LIDAR modalities, the smooth-product fusion rule is expressed by

$$\mathscr{F}_{Prod} = \frac{\prod_{i=1}^{m}(\rho_i + \alpha)}{\prod_{i=1}^{m}(\rho_i + \alpha) + \prod_{i=1}^{m}(1 - \rho_i + \alpha)} \tag{3}$$

where α is the additive smoothing factor and $m = 2$ (CNN-models based on DM and RM). The influence of α on the smoothed scores has to be minimal in order to keep the new values of ρ consistent with its distribution. A practical range for α is in the interval $(0, 0.1]$. In our experiments, we considered $\alpha = 0.05$.

VII. Experiments and Results

All results were analyzed using F-score performance measure and ROC curves, allowing a more detailed and accurate analysis of the results. The F-score results in Tables II and III are reported as a function of the spatial-filter size *i.e.*, the mask size. The F-scores values were obtained considering a threshold of 0.5. The number of pedestrian and non-pedestrian examples is unbalanced, as shown in Table I, thus, F-score is here considered because it is a suitable performance measure for unbalanced cases. Based on the classification performance results using single channels CNN (a CNN for DMs and another for RMs), as shown in Tables II and III, we chose the DM and RM maps as generated by 9×9 mask size and bilateral filter (BF).

TABLE II
F-SCORE RESULTS ON DMS FOR INCREASING SPATIAL-FILTER SIZE AND FOR DIFFERENT INTERPOLATION TECHNIQUES.

Filter size	Ave	BF	IDW	Max	Min
9×9	0.85	**0.86**	0.83	0.85	0.83
11×11	0.84	0.86	0.85	0.83	0.84
13×13	0.85	0.85	0.85	0.85	0.82
15×15	0.85	0.86	0.83	0.84	0.83

TABLE III
F-SCORE RESULTS ON REFLECTANCE MAPS (RMS).

Filter size	Ave	BF	IDW	Max	Min
9×9	0.87	**0.90**	0.89	0.84	0.83
11×11	0.85	0.88	0.89	0.85	0.84
13×13	0.87	0.87	0.87	0.83	0.82
15×15	0.84	0.80	0.83	0.72	0.83

TABLE IV
RESULTS USING THE FUSION STRATEGIES.

	Ave	Max	Min	Prod	CNN2ch
F-score	0.91	0.89	0.87	0.91	0.89

Figure 6 shows the ROC curves, calculated on the testing set, for the CNN models using depth map (DM) and reflectance map (RM) with BF's spatial-window 9×9. In addition, optimal operating points for threshold equal to 0.5 are shown in the curves and the values are indicated in the legend - designated by the superscript $(\circ)$, followed by $[FP, TP]$. The curves are zoomed, both for true positive rate (TP) and false positive rate (FP), for better visualization.

Figures 7 and 8 show the ROC curves for the fusion strategies. The results for the deterministic fusion rules, as described in Section VI, are shown in Fig. 7 for the early fusion (using a 2-channel CNN), while the testing results for the late fusion is given in Fig. 8. In Table IV we have the F-score values for the late and early - designated by CNN2ch - fusion strategies.

VIII. CONCLUSION AND REMARKS

This paper presented a study on pedestrian classification based on deep-CNN and data-fusion strategies. Classification performance evaluation is based on ROC and F-scores calculated in the testing-set of a LIDAR classification dataset generated from the KITTI Object dataset. KITTI provides labels for a variety of categories, namely pedestrians, cyclists, cars, vans, trains and people sitting. Therefore, we created a 'binary classification' dataset consisting of pedestrian and non-pedestrian (all remaining categories). KITTI also provides the corresponding LIDAR scans, which contains the 3D coordinate points as well as the reflectance data. For the LIDAR data, and by using spatial filtering, we calculated depth (DM) and reflectance (RM) maps to allow a direct implementation of CNN-based models.

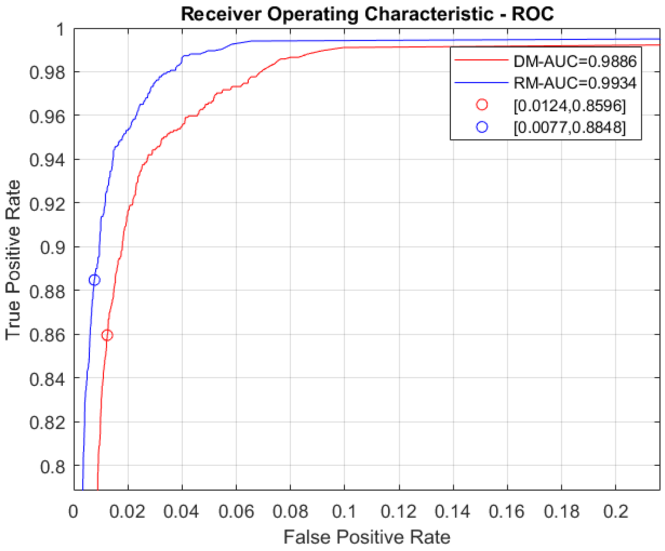

Fig. 6. ROC curves, on the testing set, using CNN on DM and RM. AUC stands for the Area Under the Curve.

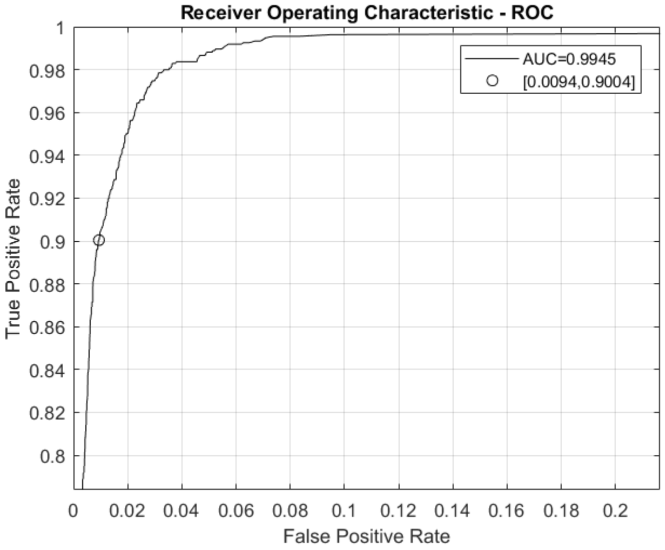

Fig. 7. ROC curves for early-fusion scheme, which is based on a 2-channels CNN (denoted by CNN2ch) using RM and DM maps.

The performances of the individually classified datasets, measured by the F-score, achieved good results, around 85%. When we considered fusions strategies, the results were better, around 90%. In addition to the F-scores measures, the results of our Bilateral filter implementation had areas under ROC close to 99%. Observing the results of F-scores measurements and ROC curves, the study carried out in this paper shows the efficiency of pedestrian classification using depth and

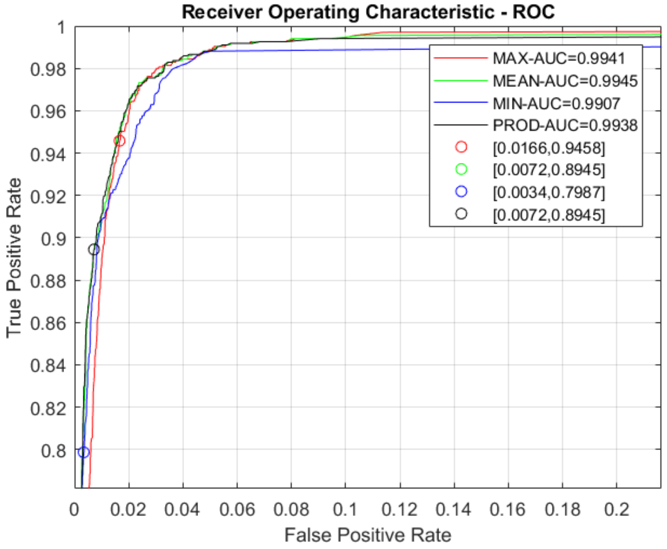

Fig. 8. ROC curves for the late-fusion rules. Results using BF spatial-filtering with mask size of 9 × 9.

reflectance maps from single LIDAR scans.

Acknowledgments

This work has been partially supported by MATIS - "Materiais e Tecnologias Industriais Sustentáveis" - reference CENTRO-01-0145-FEDER-000014, Portugal; and "AUTOCITS: Regulation Study for Interoperability in the Adoption of Autonomous Driving in European Urban Nodes" - Action number 2015-EU-TM-0243-S, co-financed by the European Union (INEA-CEF); and by the Institute of System and Robotics (ISR - Univ. Coimbra); and also by the Federal Institute of Espírito Santo-Brazil.

References

[1] P. E. Ross, "Uber robocar kills pedestrian, despite presence of safety driver," iEEE Spectrum; accessed: March, 2018. [Online]. Available: https://spectrum.ieee.org/cars-that-think/transportation/self-driving/uber-robocar-kills-pedestrian-despite-presence-of-safety-driver

[2] S. Munder and D. Gavrila, "An experimental study on pedestrian classification," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 28, no. 11, pp. 1863–1868, Dec 2009.

[3] S. Aly, "Partially occluded pedestrian classification using histogram of oriented gradients and local weighted linear kernel support vector machine," *IET Computer Vision*, vol. 8, no. 6, pp. 620–628, Dec 2014.

[4] K. Li, X. Wang, Y. Xu, and J. Wang, "Density enhancement-based long-range pedestrian detection using 3-d range data," *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 5, pp. 1368–1380, May 2016.

[5] T. E. Wu, C. C. Tsai, and J. I. Guo, "Lidar/camera sensor fusion technology for pedestrian detection," in *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, Dec 2017, pp. 1675–1678.

[6] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *International Journal of Robotics Research (IJRR)*, vol. 32, no. 11, pp. 1231–1237, sep 2013.

[7] C. Premebida, L. Garrote, A. Asvadi, A. P. Ribeiro, and U. Nunes, "High-resolution lidar-based depth mapping using bilateral filter," in *2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)*, Rio de Janeiro, Brazil, Nov 2016, pp. 2469–2474.

[8] A. Miron, A. Rogozan, S. Ainouz, A. Bensrhair, and A. Broggi, "An evaluation of the pedestrian classification in a multi-domain multi-modality setup," *Sensors*, vol. 15, no. 6, pp. 13 851–13 873, 2015.

[9] J. Schlosser, C. K. Chow, and Z. Kira, "Fusing lidar and images for pedestrian detection using convolutional neural networks," in *2016 IEEE International Conference on Robotics and Automation (ICRA)*, May 2016, pp. 2198–2205.

[10] D. O. Pop, A. Rogozan, F. Nashashibi, and A. Bensrhair, "Incremental cross-modality deep learning for pedestrian recognition," in *2017 IEEE Intelligent Vehicles Symposium (IV)*, June 2017, pp. 523–528.

[11] J. Wagner, V. Fischer, M. Herman, and S. Behnke, "Multispectral pedestrian detection using deep fusion convolutional neural networks," in *24th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*, Apr 2016, pp. 509–514.

[12] A. Asvadi, L. Garrote, C. Premebida, and U. Nunes, "DepthCN: Vehicle detection using 3d-lidar and convnet," in *2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)*, Yokohama, Japan, Oct 2017, pp. 1–6.

[13] A. Asvadi, L. Garrote, C. Premebida, P. Peixoto, and U. J. Nunes, "Real-time deep convnet-based vehicle detection using 3d-lidar reflection intensity data," in *ROBOT 2017: Third Iberian Robotics Conference*. Springer International Publishing, 2018, pp. 475–486.

[14] W. Ouyang, H. Zhou, H. Li, Q. Li, J. Yan, and X. Wang, "Jointly learning deep features, deformable parts, occlusion and classification for pedestrian detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, Aug 2018.

[15] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov 1998.

[16] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proceedings of the 25th International Conference on Neural Information Processing Systems*. Lake Tahoe, Nevada, USA: Curran Associates, Inc., 2012.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas, NV, USA: IEEE, June 2016, pp. 770–778.

[18] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Boston, MA, USA: IEEE, June 2015, pp. 1–9.

[19] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations (ICLR)*. San Diego, California, USA: ICLR, May 2014.

[20] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Computer Vision – ECCV 2014*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds., vol. 8689. Cham: Springer International Publishing, 2014, pp. 818–833.

[21] X. Chen, H. Ma, J. Wan, B. Li, and T. Xia, "Multi-view 3d object detection network for autonomous driving," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017, pp. 6526–6534.