# Pedestrian Recognition through Different Cross-Modality Deep Learning Methods

Danut Ovidiu Pop, Alexandrina Rogozan, Fawzi Nashashibi, Abdelaziz Bensrhair

# Pedestrian Recognition through Different Cross-Modality Deep Learning Methods

Dănuţ Ovidiu Pop[1], Alexandrina Rogozan[2], Fawzi Nashashibi[3] and Abdelaziz Bensrhair[4]

***Abstract*— A wide variety of approaches have been proposed for pedestrian detection in the last decay and it still remains an open challenge due to its outstanding importance in the field of automotive. In recent years, deep learning classification methods, in particular convolutional neural networks, combined with multi-modality images applied on different fusion schemes have achieved great performances in computer vision tasks. For the pedestrian recognition task, the late-fusion scheme outperforms the early and intermediate integration of modalities. In this paper, we focus on improving and optimizing the late-fusion scheme for pedestrian classification on the Daimler stereo vision data set. We propose different training methods based on Cross-Modality deep learning of Convolutional Neural Networks (CNNs): (1) a correlated model, (2) an incremental model and, (3) a particular cross-modality model, where each CNN is trained on one modality, but tested on a different one. The experiments show that the incremental cross-modality deep learning of CNNs achieves the best performances. It improves the classification performances not only for each modality classifier, but also for the multi-modality late-fusion scheme. The particular cross-modality model is a promising idea for automated annotation of modality images with a classifier trained on a different modality and/or for cross-dataset training.**

## I. INTRODUCTION

Pedestrian detection is one of the greatest importance issues in the domain of object recognition and computer vision. It is an essential problem in special in the field of the surveillance and automotive safety [1], where an efficient Advanced Driver Assistance System (ADAS) for pedestrian detection is required to diminish the number of accidents and fatal injuries[1]. The ADAS systems usually include entities for capture the road environment (multi-modality sensors and/or camera networks) followed by processing components to extract pertinent features from signals/ images which are then classified by recognition components.

In 2015, the ABI Research points out that Mercedes-Benz, Volvo and BMW are the leaders for the car enhancing ADAS systems. The BMW cars have been fitted with a Driver Assistance package for Pedestrian Warning, based on infrared night-vision and monocular vision cameras since 2013. The Mercedes system monitors the environments from the front of the vehicle using stereo vision cameras with long, medium and short-range radars. An Advanced Radar Sensor able to detect both objects and pedestrians, at a distance of up to 170 was proposed by the Continental company in 2016 and it is standard for VW Tiguan.

These existing ADAS systems still do not return a suitable result in all traffic situations, especially in a crowded urban environment where they are not able to detect all partially occluded pedestrians and they do not function efficiently in all extreme weather conditions. In the current traffic situations, require an efficient, robust and reliable ADAS system that does not compromise the safety of the people involved. We believe it is necessary to improve the classification component of an ADAS system to be able to discriminate between the obstacle type (pedestrian, cyclist, child, old person) in order to adapt the car driver system behavior according to the estimated risk level.

In recent research studies, deep learning neural networks including convolutional neural networks (CNNs), like LeNet, AlexNet, GoogLeNet, have usually led to improvement in classification performance [2], [3], [4]. The significant performances were achieved by combining the deep learning classification methods with multi-modality images within different fusion schemes. The drawback of those models is that they need a large amount of annotated data for each modality. It usually happens that one has not (enough) annotated data in one modality compared with other modalities.

The paper contains two independent questions: The first part studies three methods of integrating different image modalities (Intensity, Depth, Optical Flow) for the purpose of improving pedestrians detection. The second part studies how learning representations from one modality would enable prediction for other modalities, which they term as cross modality. This paper proposes to study these questions through various experiments based on the Daimler stereo vision data set.

The paper is organized as follows: Section 2 briefly presents existing approaches from the literature. Section 3 describes the architecture of our pedestrian classifier and the methods we proposed based on Cross-Modality deep learning of CNNs. Section 4 is convened with the experiments and their results on the Daimler dataset. Finally, Section 5 presents our conclusion and future work.

[1]Dănuţ Ovidiu Pop is a PhD student at RITS Team, INRIA Paris, 2 Rue Simone IFF, 75012 Paris, France in collaboration with Normandie Univ, INSA Rouen, LITIS, 76000 Rouen, France and Department of Computer Science, Babeş-Bolyai University, 7- 9 Universitatii street, 400084 Cluj-Napoca, Romania. danut-ovidiu.pop@inria.fr

[2]Dr. Alexandrina Rogozan is Associate Professor at Normandie Univ, INSA Rouen, LITIS, 76000 Rouen, France. alexandrina.rogozan@insa-rouen.fr

[3]Dr. Fawzi Nashashibi is the head of RITS Team at INRIA Paris, 2 Rue Simone IFF, 75012 Paris, France. fawzi.nashashibi@inria.fr

[4]Dr. Abdelaziz Bensrhair is Professor at Normandie Univ, INSA Rouen, LITIS, 76000 Rouen, France. abdelaziz.bensrhair@insa-rouen.fr

[1]According to European Commission statistics published in 2016, the number of pedestrians injured in road accidents in 2014 was 1,419,800 and there were 25,900 fatalities

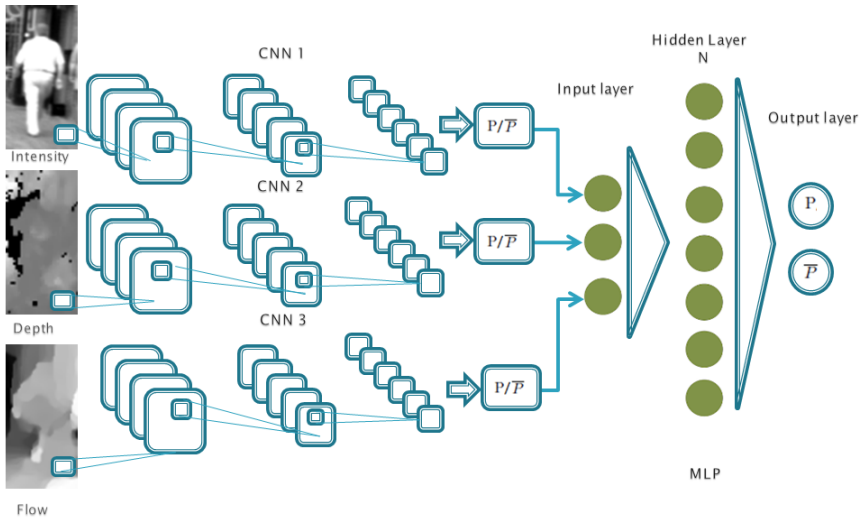



Fig. 1. Late Fusion of Intensity, Depth and Flow Modalities

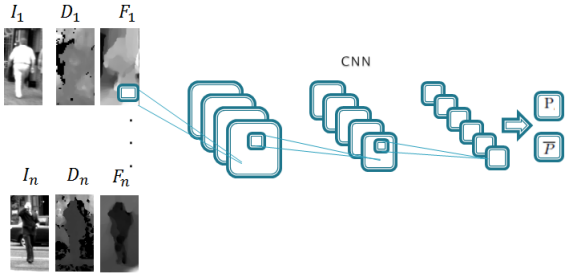



Fig. 2. Correlated Cross-Modality Training

## II. PREVIOUS WORK

The pedestrian detection is one of the most significant issues in computer vision research and objects recognition. Over the last decade, this problem was deeper investigated consist in the development of classification methods using a combination of features such as Integral Channel Features, Histograms of Oriented Gradients (HOG), Local Binary Patterns (LBP), Scale Invariant Feature Transform (SIFT), among others, followed by a trainable classifier such as a Support Vector Machine (SVM), Multilayer Perceptrons (MLP), boosted classifiers or random forests [5], [6].

A mixture-of-experts framework performed with HOG, LBP features and MLP or linear SVM classifiers was developed in [7]. An end-to-end CNN features learn approach was presented in [8]. This experiment concentrated on the detection of small scale pedestrians on the Caltech data set. A combination of three CNNs to detect pedestrians at distinct scales was proposed on the same monocular vision data set [9]. A cascade Aggregated Channel Features detector is utilized in [10] to engender candidate pedestrian windows followed by a CNN-based classifier for checking purposes on monocular Caltech and stereo ETH data sets. Two CNN-based fusion methods of visible and thermal images on the KAIST multi-spectral pedestrian data set were presented in [11]. The first approach joins the data of these modalities at the pixel level (early fusion), while the second one utilizes separate sub-networks to generate a feature representation for each modality before classification (intermediate fusion). The authors demonstrated that the intermediate fusion outperforms the early fusion.

The performance of the early fusion and late fusion models on the Daimler stereo vision data set were compared in our previous work [12]. The early fusion model was built by concatenating three image modalities (intensity, depth and optical flow) to feed a unique CNN. The late fusion model consists in fusing the outputs scores (the class probability estimate) of three independent CNNs, trained on intensity, depth and optical flow images, by an SVM classifier. Our experiment confirmed that, the early-fusion model is less efficient than the late-fusion model. Furthermore, the early fusion is less robust than the late fusion, since it needs strong image calibration and synchronization.

In the literature, for the intermediate and late fusion architectures, the training is implemented independently on each modality, with annotated images provided exclusively from that modality. To the best of our knowledge, any study has been implemented on cross-modality training for pedestrian recognition, but only on cross-dataset training. In [13], the authors proposed an incremental cross-dataset learning algorithm for the pedestrian detection problem. A synthetic dataset (Virtual Pedestrian dataset [14]) is utilized for basically training and two distinctions real-world datasets (KITTI Vision Benchmark Suite and the Daimler Mono Pedestrian Detection Benchmark) for fine-tuning and evaluation.

The aim of this paper is to improve the late-fusion training by using a cross-modality approach. We propose different training methods based on Cross-Modality deep learning of CNNs: (1) a correlated model where a unique CNN is trained with Intensity, Depth and respectively Flow images for each frame (2) an incremental model where a CNN is trained

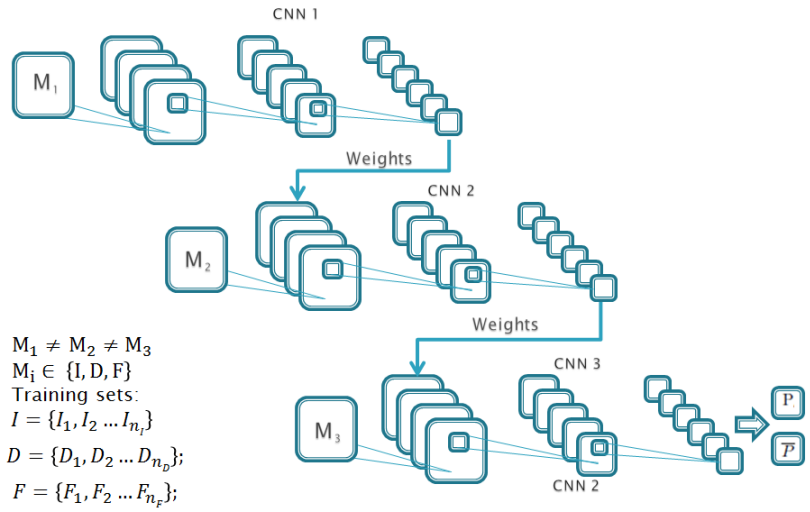



Fig. 3. Incremental Cross-Modality Training

with the first modality images frames, then a second CNN, initialized by transfer learning on the first CNN, is trained on the second modality image frames, and finally a third CNN initialized on the second CNN, is trained on the last modality images frames, (3) a particular cross-modality model where a CNN is trained on one modality, but tested on a different one. The CNNs trained with (1),(2) and (3) schemes are tested on Intensity, Depth and respectively Flow images. We will prove that the incremental cross-modality method is effective for the training of modality classifier not only with images from that modality, but also with images from other modalities among Intensity, Depth and Optical Flow. The particular cross-modality model (3) is a promising idea for automated annotation of modality images with a classifier trained on a different modality and/or for cross-dataset training. All these models are presented in the following section.

## III. THE PROPOSED ARCHITECTURES

In this paper, we propose fusing stereo-vision information between three modalities: Intensity (I), Depth (D) and Optical Flow (F). We investigate the late-fusion architecture using three different methods for the training of the CNN-based classifiers: a classical intra-modality approach and three different methods for a cross-modality approach. The three methods represent how different image modalities are used to train CNNs : (1) The baseline method is where each image modality is parallelly fed into independent CNNs and the classifier probabilistic scores from these CNNs are fused using a multilayer perceptron (2) The incremental method is where samples from each modality are sequentially fed to a single CNN (3) The correlation method is where set of images from the same frame having different image modalities are the input of single CNN.

### A. Our baseline late fusion architecture

We propose a late-fusion architecture (see Fig 1) where an MLP is used to discriminate between pedestrians (P) and non-pedestrians ($\overline{P}$) on the classification results (it combines the output scores of all classifiers) of three modality CNNs. Each CNN is exclusively trained with images from the same modality (among intensity, depth and optical flow) and then tested on that modality images.

Each modality CNN is based on the LeNet architecture which consists of 7 layers, excepting the input layer, 2 convolutional layers, 2 pooling layers, 2 inner product (IP) layers and one rectified linear unit (ReLU) layer. We use 20 filters with one stride for the first convolutional layer followed by 50 filters with one stride for the second one. We use two IP layers with 500 neurons for the first IP layer and 2 neurons for the second IP layer. The final layer returns the classifier probabilistic scores from CNN's and after that, an MLP fused the classifier probabilistic scores for obtaining the final decision of the classifier system: P or $\overline{P}$.

### B. Correlated cross-modality training

We propose a Correlated Cross-Modality (CCM) approach where a unique CNN is learned with the same image frames, but provided in different modalities, Intensity $I_i$, Depth $D_i$ and Flow $F_i$ with i=$\overline{1,n}$ (see Fig. 2). The CNN model is validated in two different ways: on a multi-modal validation set (stack of images from the same frame for different image modalities) and respectively on a single modality set (see table I ). The learning and validation sets are disjointed.

We believe that the drawback of the correlated cross-modality training is that it compels one to use a unique CNN model. This is a too strong constraint, if different modalities could improve the learning process with different CNN architectures and/or with different settings (i.e. learning algorithms and learning rates).

### C. Incremental cross-modality training

Our experiments show that modality CNNs provide the best results with different architectures and settings [15].

TABLE I

COMPARISON OF CORRELATED VS INCREMENTAL CROSS-MODALITY TRAINING MODELS

| Approach | Learned on | Validated on | Tested on | TPR | FPR | ACC |
|---|---|---|---|---|---|---|
| CoCM | $\text{Intensity}_i$+$\text{Depth}_i$+$\text{Flow}_i$ i=$\overline{1,n}$ | $\text{Intensity}_j$+$\text{Depth}_j$+$\text{Flow}_j$ j=$\overline{1,m}$ | Intensity | 0.972 | 0.0713 | 94.54% |
| | | $\text{Intensity}_j$+$\text{Depth}_j$+$\text{Flow}_j$ j=$\overline{1,m}$ | Depth | 0.9089 | 0.1809 | 85.39% |
| | | $\text{Intensity}_j$+$\text{Depth}_j$+$\text{Flow}_j$ j=$\overline{1,m}$ | Flow | 0.9161 | 0.1386 | 88.26% |
| | | Intensity | Intensity | 0.972 | 0.0737 | 94.4% |
| | | Depth | Depth | 0.9112 | 0.0172 | 86.06% |
| | | Flow | Flow | 0.9115 | 0.152 | 87.38% |
| InCM | $\text{Depth}_i$,$\text{Flow}_i$,$\text{Intensity}_i$ i=$\overline{1,n}$ | $\text{Depth}_j$,$\text{Flow}_j$,$\text{Intensity}_j$ j=$\overline{1,m}$ | Intensity | 0.9619 | 0.029 | 96.7% |
| | $\text{Intensity}_i$,$\text{Flow}_i$,$\text{Depth}_i$ i=$\overline{1,n}$ | $\text{Intensity}_j$,$\text{Flow}_j$,$\text{Depth}_j$ j=$\overline{1,m}$ | Depth | 0.8764 | 0.095 | 89.39% |
| | $\text{Depth}_i$,$\text{Intensity}_i$,$\text{Flow}_i$ i=$\overline{1,n}$ | $\text{Depth}_j$,$\text{Intensity}_j$,$\text{Flow}_j$ j=$\overline{1,m}$ | Flow | 0.9436 | 0.056 | 94.34% |

TABLE II

COMPARISON OF CLASSICAL TRAINING (TRAINING AND TESTING ON THE SAME MODALITY) VS PARTICULAR CROSS-MODALITY (TRAINING ON ONE MODALITY AND TESTED ON A DIFFERENT ONE)

| Trained on | Tested on | ACC |
|---|---|---|
| Intensity | Intensity | 96.55% |
| Depth | Intensity | 50.51% |
| Flow | Intensity | 73.79% |
| Intensity | Depth | 58.24% |
| Depth | Depth | 89.78% |
| Flow | Depth | 54.33% |
| Intensity | Flow | 72.97% |
| Depth | Flow | 57.55% |
| Flow | Flow | 87.34% |

Therefore, we propose an Incremental Cross-Modality (ICM) training, based on a transfer learning approach (see Fig.3).

A first CNN is trained and validated with the first modality images frames, then a second CNN, initialized by transfer learning on the first CNN, is trained and validated on the second modality images frames, and finally a third CNN initialized on the second CNN, is trained and validated on the last modality images frames. Transfer learning consists of transferring the weight information from a previous CNN, that has already been learned, to a new CNN which will be trained next [16].

The advantage of this method is that its architecture is more flexible allowing for adaptive settings for each classifier (i.e. different learning algorithms and rate policies). Learning this model does not require any inter-modality correlated data, nor synchronized modality images. This could be an interesting point if the multi-modality images are unbalanced, various and not acquired with similar sensors/cameras and/or extracted from the same database. The approach can be extended to cross-dataset training.

For this approach, the question is whether the order of modality training within the previous model has any importance. We investigate different combinations and we conclude that for the classification in the Intensity modality the optimal order for training is Depth images first, followed by Flow images and finally Intensity images (D,F,I training model of I); for the classification in the Flow modality the optimal order for training is Depth images first, followed by Intensity images and finally Flow images (D,I,F training model of F), and respectively for the classification in the Depth modality the optimal order for training is Intensity images first, followed by Flow images and finally Depth images (I,F,D for training of D) (see Table I).

## IV. EXPERIMENTS AND RESULTS

The training and testing were carried out on Daimler stereo vision images of 48 x 96 px with a 12-pixel border around the pedestrian images extracted from three modalities: Intensity, Depth and optical Flow.

We use 84577 samples for training, 75% of which are used for learning, 25% for validation (optimization of CNN's hyperparameters) and 41834 for testing. The training set contains:

- 52112 samples of pedestrians
- 32465 samples of non pedestrians

The testing set contains:

- 25608 samples of pedestrians
- 16235 samples of non pedestrians

The experiments are performed in the Caffe deep neural network framework. The performances are measured by the Accuracy (ACC) and using the Receiver Operating Characteristics (ROC) curve created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. The complexity of the classification system is also investigated by the area under the curve (AUC).

### *A. Comparison of uni-modal classifiers with classical training*

We start by comparing for each modality images the classification performances with LeNet architecture with different learning algorithms: Stochastic Gradient Descent (SGD), Adaptive Gradient (ADAGRAD), RMSPROP, ADADELTA and learning rate polices: Fixed (FIX), Exponential (EXP), Step Down (STEP), Polynomial Decay (POLY), Sigmoid (SIG), Multi-Step (MS) and Inverse Decay (INV) [15]. Each modality classifier is exclusively trained and validated with

TABLE III
PERFORMANCE WITH LATE FUSION ON DAIMLER TESTING SET

| Late-fusion | Trained on | TPR | FPR | AUC |
|---|---|---|---|---|
| Classical training | Same settings (RMSPROP - POLY) | 0.951 | 0.0109 | 97.04% |
| Incremental Cross Modality training | Same settings (RMSPROP - POLY) | 0.953 | 0.0092 | 97.20% |
| | Specific modality settings (RMSPROP - POLY, SGD - EXP, ADADELTA - FIX) | 0.973 | 0.023 | **97.47%** |

images of its own modality. For the Intensity modality the best performance (ACC = 96.55%) was achieved with the LeNet architecture using the RMSPROP [2] algorithm learning with POLY rate policy. The best performances are obtained in Depth images with SGD and EXP settings (ACC = 89.78%) and respectively in Flow images with ADADELTA and FIX settings (ACC = 87.34%). Therefore, different modalities need different learning algorithms and rate policies for effective performances. All the CNNs were optimized on the training set through 29760 epochs and 0.01 learning rate.

### *B. Comparison of uni-modal classifiers with cross-modality training*

*1) Correlated cross-modality:* Since the RMSPROP with POLY settings achieved the best performance on the Intensity modality, we decided to use these settings to train the correlated cross-modality models (CCM). The CNN models are validated following two different approaches on the multi-modality union data set or on specific modality ones (see Table I). The first approach gives better results. This correlated cross-modality training outperforms the classical intra-modality training, but only on the testing set for Flow modality by achieving an accuracy of ACC=88,26 %. This may be explained by the fact that, with more complex training data, the breadth and depth of the network should be increased. However according to [17], the complexity would be limited by the computing resources, which would thus hinder performance (see Table I). The CNNs belonging to the correlated cross-modality approach are trained with three times more training epochs (89220 epochs) than the classical ones (29760 epochs) for the same learning rate (0.01).

*2) Incremental cross-modality:* For the training with the incremental cross-modality method (ICM), we use RMSPROP with POLY settings for all three CNNs through 29760 epochs with the same learning rate (0.01). The results, given in Table I) are better than those achieved with both the classical training [15] and the correlated cross-modality training for the same settings.

*3) Particular cross-modality:* We also tested particular cross-modality models where each CNN-based classifier is trained on one modality, but tested on a different one (see Table II). The best performance for this approach is achieved on Intensity images when trained on Flow images (ACC = 73.79%), on Depth images when trained on Intensity images (ACC = 58.24%), and respectively on Flow images when trained on Intensity images (ACC = 72.97%). The performances are below those obtained when the training and testing are performed on the same modality, but the results are quite interesting ones. This could be a promising idea for semi-automatic annotation of modality images with a classifier trained with other modality data.

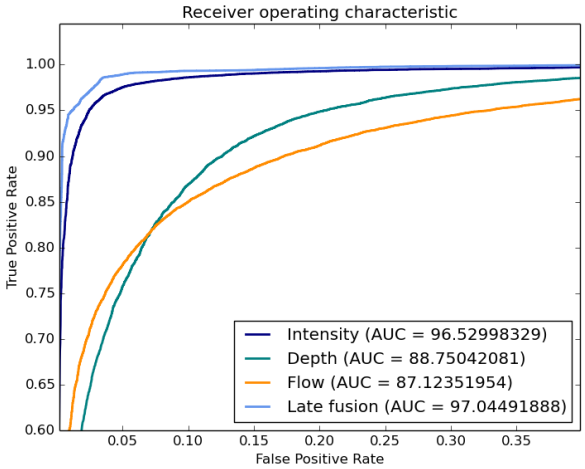
[2]Tieleman, T. and Hinton, G.,Lecture 6.5—RmsProp: Divide the gradient by a running average of its recent magnitude, COURSERA: Neural Networks for Machine Learning, 2012

### *C. Late-fusion with classical vs cross-modality training*

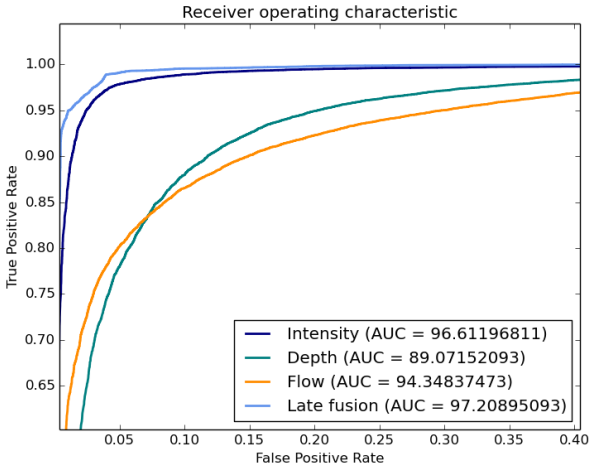
This approach was investigated in two ways. First, we decided to use the same setting to train all the CNNs within the incremental cross-modality model. We have chosen the RMSPROP with POLY setting, since it achieved the best performance on the Intensity modality. Second, we decided to use the best specific settings to train the CNNs through the incremental cross-modality model, and consequently the RMSPROP with POLY setting for the Intensity modality, the SGD with EXP setting for Depth modality and ADADELTA with FIX settings for the Flow modality. In Table III we show that the performance obtained with incremental cross-modality using the best specific modality settings (RMSPROP with POLY for Intensity, SGD with EXP for Depth and ADADELTA with FIX for Optical Flow) are better than those obtained with the same setting whatever they are within a classical training or an incremental cross-modality one. The incremental cross-modality training we propose is an efficient solution not only within the single modality classifiers but also within the late-fusion scheme. These performances are also shown in the ROC curves (see Fig 4).

## V. CONCLUSION

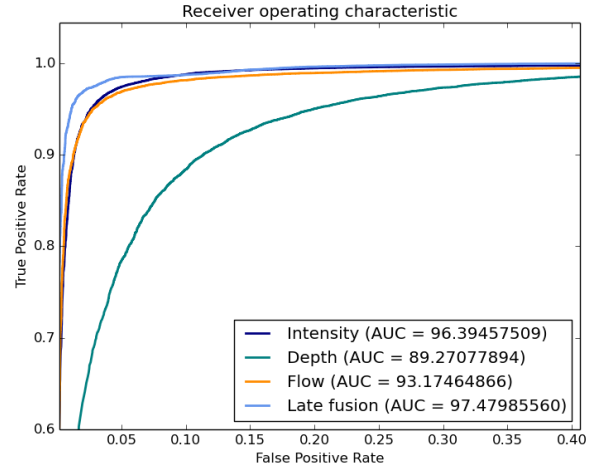
In this paper, we proposed differently cross-modality training approaches for late-fusion architectures to improve pedestrian recognition. The incremental cross-modality approach outperforms the correlated cross-modality approach. The incremental method also improves the classification performances compared to a classical training of uni-modal CNNs through late-fusion schemes on the Daimler data set. We believe that the incremental approach is the promising one. Future work will be concerned with improving the incremental cross-modality training by extending the method to cross datasets training. We are going to develop the late fusion architecture on the particular cross-modality with improving this model by using optimal settings for different training modality sets. Moreover, we intend to apply this presented architecture to classify any other road objects and




(a) Classical cross-modality training with RMSPROP-POLY setting




(b) Incremental cross-modality training with RMSPROP-POLY setting




(c) Incremental cross-modality training using best modality specific setting

Fig. 4. ROC classification performance on Daimler testing data set.

participants, such as vehicles, cyclists, traffic signs and traffic lights.

## VI. ACKNOWLEDGEMENTS

The research for this paper was financially supported by the Normandy Region and Inria Paris.

## REFERENCES

[1] Pierre Sermanet, Koray Kavukcuoglu, Soumith Chintala, and Yann Lecun. Pedestrian detection with unsupervised multi-stage feature learning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2013.

[2] Jan Hosang, Mohamed Omran, Rodrigo Benenson, and Bernt Schiele. Taking a deeper look at pedestrians. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[3] H. Fukui, T. Yamashita, Y. Yamauchi, H. Fujiyoshi, and H. Murase. Pedestrian detection based on deep convolutional neural network with ensemble inference network. In *2015 IEEE Intelligent Vehicles Symposium (IV)*, pages 223–228, June 2015.

[4] Anelia Angelova, Alex Krizhevsky, and Vincent Vanhoucke. Pedestrian detection with a large-field-of-view deep network. In *IEEE International Conference on Robotics and Automation, ICRA 2015, Seattle, WA, USA, 26-30 May, 2015*, pages 704–711, 2015.

[5] Rodrigo Benenson, Mohamed Omran, Jan Hosang, and Bernt Schiele. *Ten Years of Pedestrian Detection, What Have We Learned?*, pages 613–627. Springer International Publishing, Cham, 2015.

[6] Piotr Dollar, Christian Wojek, Bernt Schiele, and Pietro Perona. Pedestrian detection: An evaluation of the state of the art. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(4):743–761, April 2012.

[7] M. Enzweiler and D. M. Gavrila. A multilevel mixture-of-experts framework for pedestrian classification. *IEEE Transactions on Image Processing*, 20(10):2967–2979, Oct 2011.

[8] R. Bunel, F. Davoine, and Philippe Xu. Detection of pedestrians at far distance. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2326–2331, May 2016.

[9] M. Eisenbach, D. Seichter, T. Wengefeld, and H. M. Gross. Cooperative multi-scale convolutional neural networks for person detection. In *2016 International Joint Conference on Neural Networks (IJCNN)*, pages 267–276, July 2016.

[10] Xiaogang Chen, Pengxu Wei, Wei Ke, Qixiang Ye, and Jianbin Jiao. *Pedestrian Detection with Deep Convolutional Neural Network*, pages 354–365. Springer International Publishing, Cham, 2015.

[11] Jörg Wagner, Volker Fischer, Michael Herman, and Sven Behnke. Multispectral pedestrian detection using deep fusion convolutional neural networks. In *24th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*, pages 509–514, April 2016.

[12] Dănuţ Ovidiu Pop, Alexandrina Rogozan, Fawzi Nashashibi, and Abdelaziz Bensrhair. Fusion of stereo vision for pedestrian recognition using convolutional neural networks. *In 25th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*, pages 47–52, April 2017.

[13] C. Karaoguz and A. Gepperth. Incremental learning for bootstrapping object classifier models. In *2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)*, pages 1242–1248, Nov 2016.

[14] David Vazquez, Antonio M. Lopez, Javier Marin, Daniel Ponsa, and David Geronimo. Virtual and real world adaptation for pedestrian detection. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 36(4):797–809, 2014.

[15] Dănuţ Ovidiu Pop, Alexandrina Rogozan, Fawzi Nashashibi, and Abdelaziz Bensrhair. Incremental cross-modality deep learning for pedestrian recognition. *Proceedings of the IEEE Intelligent Vehicles Symposium*, 2017.

[16] S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, Oct 2010.

[17] Yann Lecun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. In *Proceedings of the IEEE*, pages 2278–2324, 1998.