Improving Pedestrian Recognition using Incremental Cross Modality Deep Learning

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

1 - INRIA Paris - RITS Team
Paris - France

2 - Normandie Univ - INSA Rouen, LITIS
Rouen - France

3 - Babeş-Bolyai University - Department of Computer Science
Cluj-Napoca - Romania

Abstract. Late fusion schemes with deep learning classification patterns set up with multi-modality images have an essential role in pedestrian protection systems since they have achieved prominent results in the pedestrian recognition task. In this paper, the late fusion scheme merged with Convolutional Neural Networks (CNN) is investigated for pedestrian recognition based on the Daimler stereo vision data sets. An independent CNN-based classifier for each imaging modality (Intensity, Depth, and Optical Flow) is handled before the fusion of its probabilistic output scores with a Multi-Layer Perceptron which provides the recognition decision. In this paper, we set out to prove that the incremental cross-modality deep learning approach enhances pedestrian recognition performances. It also outperforms state-of-the-art pedestrian classifiers on the Daimler stereo-vision data sets.

1 Introduction

The detection and classification of pedestrians have attracted a great deal of attention from researchers due to its vast applicability for autonomous vehicles and driver assistance systems (ADAS). Fatigue, discomfort, alcohol consumption, bad visibility or pedestrians' illegal and/or unsafe behavior, are the most frequent human errors which cause traffic accidents. These collisions could be effectively reduced if such human errors were eliminated employing a Pedestrian Detection System (PDS), using a classifier component. This issue has been widely investigated, but it still remains an open challenge because PDS progress is hindered by the difficulty of detecting all partially occluded pedestrians and the problem of operating efficiently in severe weather conditions.

The objective of the work described in this paper is to analyze how an incremental cross-modality deep learning approach enhances pedestrian recognition performance by learning a CNN-based classifier using a combination of Intensity (I), Depth (D) and Optical Flow (F) modalities on the Daimler stereo vision data sets. To achieve this aim, we develop the following methodology based on two CNNs: AlexNet for its incontestable impact on machine learning due to a

good balance between performance and compact architecture, and VGG-16 because of is great performance obtained with vast a architecture commonly used in pedestrian detection.

- Learn the AlexNet and VGG-16 using the default CNNs setting with the Classical Learning Method[1] (CLM) and respectively with the Incremental Cross Modality Deep Learning (INCML) method;
- Optimize the CNNs hyper-parameters (convolution stride, kernel size, convolution number of outputs, weights of the fully connected layers) for the CLM and for the INCML method respectively;
- Implement the late fusion scheme with Multi-Layer Perceptron (MLP) for both classical and incremental methods considered above.
- Compare our learning patterns with the state-of-the-art MoE framework proposed in [1, 2] and a deep Boltzmann-Machine [3]

2 Related work

The pedestrian detection issue has attracted significant interest over the last decade, resulting in a wide variety of approaches that can be classified in two categories: handcrafted features models such as Histograms of Oriented Gradients (HOG) [1, 2], Scale Invariant Feature Transform, Integral Channel Features, Local Binary Patterns (LBP), combined with a trainable classifier such as a Multi-Layer Perceptron (MLP) [1, 2], boosted classifiers Support Vector Machine (SVM) or random forests, and deep learning neural network models especially Convolutional Neural Networks such as AlexNet [4] and VGG-16 [5].

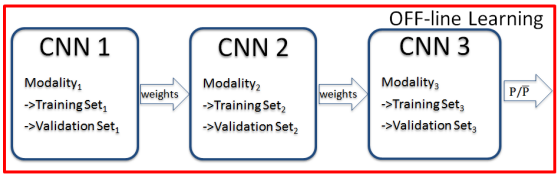
We mention only the state-of-the-art approaches delivered with the Daimler data sets since our cross-modality learning models are developed on those data sets so as to allow a fair comparison: a mixture-of-experts (MoE) framework assessed with HOG and LBP features merged with MLP [2] or linear SVM [1] classifiers are described. Those feature-based MoE models are learned using a classical learning methodology where both training and validation were done on the same modality for Intensity, Depth and Optical Flow. A deformation part-based model is combined with a deep model based on a restricted Boltzmann Machine for pedestrian detection [3]. The deformation-part component receives the scores of pedestrian body-part detectors and provides a decision hypothesis to the deep model in order to discriminate the visibility correlation between overlapping elements at multi-layers. This approach was applied not only on the Daimler data sets but also on the Caltech, ETH and CUHK data sets.

3 Architectures

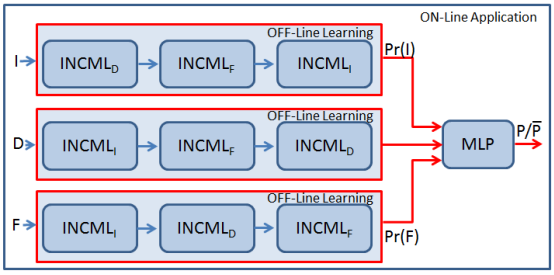
3.1 Incremental Cross-Modality Learning

The approach we proposed in [6] was used with a limited CNN (LeNet). In this article, we study how it performs with a vast CNN. It involves samples

[1]For the classical learning approach, we have trained, validated and tested each CNN with the corresponding image modality for Intensity, Depth, and Optical Flow.



(a) Incremental Cross-Modality Deep Learning.



(b) The Late Fusion Architecture.

Fig. 1: The proposed architectures.

from each modality sequentially inserted into successive CNNs using an original transfer learning technique [7]. The method investigates if the classification performance is enhanced when it uses a different image modality validation set than the training set and when data in one of the modalities is scarce (i.e., many more images in the visual spectrum than depth). The Incremental Cross-Modality learning approach implies a CNN learnt with the first images modality frames, then a second CNN, initialized by assigning the weight information from a previous CNN, is learnt on the next image modality frames, and finally a third CNN initialized on the second CNN, is learnt on the last image modality frames (see Fig. 1a).

One of the main advantages of this approach is that its architecture is flexible, allowing for adaptive specific hyper-parameters for each couple of modality/classifier (i.e. various architecture, learning algorithms and rate policies). The method can also be adapted for cross-datasets training.

We obtained the following image modalities learning order with an optimization method on the validation Daimler data set: start with Depth followed by Optical Flow then Intensity learning model for the Intensity; begin with Depth followed by Intensity then Optical Flow learning model for Optical Flow; starts with Intensity followed by Optical Flow then Depth learning model for Depth. In this paper, we also optimized the CNNs hyper-parameters for both the CLM, and the INCML approaches.

3.2 The Late-Fusion Architecture

The architecture involves three independent INCML-based classifiers [8] which are fed with a specific modality among Intensity, Depth and Optical Flow, and an MLP which discriminates between pedestrians (P) and non-pedestrians ($\overline{\text{P}}$) using probabilistic class estimates provided by each INCML classifier (see Fig. 1b). The last layer of each CNN provides the probability output scores of Intensity Pr(I), Depth Pr(D) and Optical Flow Pr(F). The MLP includes three neurons in the input layer, one hidden layer of 100 neurons, and 2 neurons in the output layer. The ReLU function with a Stochastic Gradient Descent solver and a constant learning rate of 1e-07 were used for the weights optimization.

Table 1: Comparison with the state-of-the-art on Daimler data set

Pedestrian Data Set	Method and Settings		TPR 90%	TPR=95% FPR ± CI/2
Partially Occluded (p-occ)	AlexNet	Default-CLM	0.73	0.8671 ± 0.0034%
		Default-INCLM	0.712	0.7126 ± 0.0037%
		Optim-CLM	0.137	0.2363 ± 0.0042%
		Optim-INCLM	0.105	**0.1920 ± 0.0039%**
	VGG-16	Default-CLM	0.597	0.7360 ± 0.0044%
		Default-INCLM	0.605	0.7704 ± 0.0042%
		Optim-CLM	0.447	0.6495 ± 0.0047%
		Optim-INCLM	0.457	0.6714 ± 0.0047%
	HOG/linSVM MoE [1]		0.175	0.20 ± 0.0040%
	Deep DP-BM [3]		0.216	0.25 ± 0.0043%
Non Occluded (non-occ)	AlexNet	Default-CLM	0.328	0.4465 ± 0.0048%
		Default-INCLM	0.362	0.4939 ± 0.0048%
		Optim-CLM	0.0006	**0.0011 ± 0.00031%**
		Optim-INCLM	0.0009	0.0014 ± 0.00035%
	VGG-16	Default-CLM	0.151	0.2531± 0.0042%
		Default-INCLM	0.125	0.2150± 0.0039%
		Optim-CL	0.011	0.0296± 0.0016%
		Optim-INCLM	0.0078	0.0236± 0.0015%
	HOG+LBP/MLP MoE [2]		0.0002	0.0035 ± 0.00056%
	HOG/linSVM MoE [1]		0.011	0.0302 ± 0.0016%
	Deep DP-BM [3]		0.007	0.05 ± 0.0021%

4 Experimental Setup

The learning and testing were carried out on Daimler [1] 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.

The learning set involves 84577 samples (52112 samples of non-occluded pedestrians and 32465 samples of non-pedestrians) for learning, 75% of which are used for training, 25% for validation (optimization of CNN's hyper-parameters

on the holdout validation method). The testing set includes 36768 samples of pedestrians (25608 samples of non-occluded pedestrians and 11160 samples of partially occluded pedestrians), and 16235 samples of non-pedestrians.

The learning and testing processes operate on AlexNet and VGG-16 using original (default) and optimized settings respectively. For the learning with the incremental cross-modality and classical methods, we use RMSPROP with POLY settings for all CNNs with the same learning rate (0.0001). The setting optimization consists in removing the crop of size features, reducing the number of outputs (from 96 to 20), decreasing the kernel size (from 7 to 5) and minimizing the stride (from 4 to 1 for AlexNet and from 2 to 1 for VGG-16) in the first convolution layer. We also marked down the number of outputs (from 4096 to 2048) in the last two Fully Connected (FC) layers for AlexNet and respectively in the previous three FC layers (from 4096 to 2048 in the FC6 and FC7 respectively from 1000 to 500 in the FC8) for VGG-16.

5 Analysis of results

The learning process is executed in the Caffe deep framework on GPU with Nvidia Quatro P5000. The complexity of the classification system is assessed by the False Positive Rates (FPR) using a True Positive Rate (TPR) of 95% and a Confidence Interval (CI) to prove whether one model is statistically better than another one. The results, given in Table 1, allow for the following comparisons on the Daimler data sets:

Default vs. optimized settings: The optimization method presented allows statistically significant improvement for both CML/INCML approaches and AlexNet/VGG-16 architectures up to $\Delta = 0.4925\%$

Incremental Cross-Modality Deep Learning vs. Classical Learning: With the optimized settings the results obtained with INCML are statistically better than those achieved with the CLM, but only with AlexNet on the partially occluded pedestrian Daimler data set and with VGG-16 on the non-occluded pedestrian Daimler data set.

Comparisons with the state-of-the-art classifiers provided with the Daimler data sets: The improvements obtained with optimized settings on TPR=95% based on AlexNet using the INCML method approach are statistically significant on both partially-occluded and non-occluded data sets since the confidence intervals are disjoint: ΔFPR MoE$_{p-occ}$=0.008%, ΔFPR MoE$_{non-occ}$= 0.0024%, Δ FPR DP-BM$_{p-occ}$=0.0489%, ΔFPR DP-BM$_{non-occ}$ =0.076%. On the other hand, our method does not outperform the method [1] at TPR=90%. However, [1] was only analyzed on the non-occluded pedestrian data set. Our approach was learnt on the entire data set which includes occluded and non-occluded pedestrian samples. It is to be noted that the AlexNet obtained better results than VGG-16 on the Daimler data sets, this highlighting that a huge architecture does not always achieved better results.

6 Conclusion

This paper presents evidence of how incremental-cross deep learning modality improves the pedestrian recognition system. The experiments were carried out on AlexNet, and VGG-16 using default and optimized learning settings based on an incremental cross-modality learning approach, then merged with a late fusion scheme. The incremental cross-modality approach outperforms the classical learning approach on the partially occluded pedestrian Daimler data set using AlexNet and on the non-occluded pedestrian Daimler data set using VGG-16. Indeed, this cross-modality learning method is more flexible than others since it could be used with the most suitable learning settings for each image modality. The INCML approach merged with the late fusion scheme outperforms state-of-the-art pedestrian classifiers for both non-occluded and partially-occluded pedestrian Daimler data sets. This method could enhance pedestrian recognition if the data set characteristics, CNN architecture and its hyper-parameters are adapted to the target application.

Future work will be concerned with improving and benchmarking of the incremental cross-modality learning by extending the method to cross datasets (using Caltech, KITTI, JAAD) training using multi-class detection (SSD, Faster RCNN, R-FCN) and applying the promising INCML model to the classification and detection of other road objects (traffic signs and traffic lights) and road users (vehicles, cyclists).

References

[1] Markus Enzweiler, Angela Eigenstetter, Bernt Schiele, and Dariu M. Gavrila. Multi-cue pedestrian classification with partial occlusion handling. In *CVPR*, pages 990–997. IEEE Computer Society, 2010.

[2] 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.

[3] W. Ouyang and X. Wang. A discriminative deep model for pedestrian detection with occlusion handling. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3258–3265, June 2012.

[4] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.

[5] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[6] 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)*, pages 523–528, June 2017.

[7] S.J. Pan and Q. Yang. A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010.

[8] 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.