Cross Training for Pedestrian recognition using Convolutional Neural networks

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

Cross Training for Pedestrian recognition using Convolutional Neural networks

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

[1] RITS Team, INRIA Paris, France
[2] LITIS Laboratory, INSA de Rouen, France
[3] Department of Computer Science, Babeş-Bolyai University, Cluj-Napoca, Romania

danut-ovidiu.pop@inria.fr

Abstract

In recent years, deep learning classification methods, specially Convolutional Neural Networks (CNNs), combined with multi-modality image fusion schemes have achieved remarkable performance. Hence, in this paper, we focus on improving the late-fusion scheme for pedestrian classification on the Daimler stereo vision data set. We propose cross training method in which a CNN for each independent modality (Intensity, Depth, Flow) is trained and validated on different modalities, in contrast to classical training method in which the training and validation of each CNN is on same modality. The CNN outputs are then fused by a Multi-layer Perceptron (MLP) before making the recognition decision.

Keywords

pedestrian recognition, deep learning, convolutional neural network, late-fusion, cross-training.

1 Introduction

Pedestrian detection is a key problem for surveillance, robotics applications and automotive safety where an efficient Advanced Driver Assistance System (ADAS) for pedestrian detection is needed to reduce the number of accidents and fatal injures.

A study performed by ABI Research published in 2015 shows that Mercedes-Benz, Volvo and BMW dominate the market for car enhancing ADAS systems. These existing ADAS systems still have difficulty distinguishing between human beings and nearby objects. Our work is concerned with the improvement of the classification component of a pedestrian detector. In recent research studies, deep learning neural networks including CNNs, like LeNet, AlexNet, GoogLeNet, have usually proved classification performance improvement. The drawback for those models is that they require a large amount of annotated data for each modality.

The question is could be used one modality for training and another modality for validating (standpoint one) or only the same training and validating modality (standpoint two) for improving the classification model. To our knowledge, these questions have not yet been answered for the pedestrian recognition task. This paper proposes to solve this brain-teaser through various experiments based on the Daimler stereo vision data set.

2 Previous Work

Over the last decade, the pedestrian detection has been a significant issue in computer vision research and object recognition. A wide variety of methodologies have been proposed with optimization in performance, resulting in the development of classification methods using a combination of features followed by a trainable classifier [1].

In [2] was presented a CNN to learn the features with an end-to-end approach on the Caltech data set. A combination of three CNNs to detect pedestrians at different scales was proposed on the same monocular vision data set [3]. Two CNN-based fusion methods of visible and thermal images on the KAIST multi-spectral pedestrian data set were presented in [4].

We compared in [5] the performance of the early fusion and late fusion models on the Daimler stereo vision data set. We showed the early-fusion model is less efficient than the late-fusion model. On this paper is proposing to improve the late-fusion training by using cross-training approach within a hybrid CNN-MLP framework. Each imaging modality among Intensity, Depth and Flow, is training on one image modality and validating on the other one by an independent CNN. The CNN outputs are then fused by a MLP to improve the recognition decision.

3 The Proposed Architectures

In this paper, we propose fusing stereo-vision information between three modalities: Intensity (I), Depth (D) and Flow(F). 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 (the class probability estimate) of three modality CNNs. Each CNN is exclusively trained with images from the same modality (among intensity, depth and flow) and then tested on that modality images. We compare the classical-training method where each imaging modality among Intensity, Depth and Flow, is classified by an independent CNN and cross-training method where each imaging modality is training exclusively on one modality and is validating on the other modality (see Table 1).

Each modality CNN is based on the LeNet architecture. 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 final decision of the classifier system: P or $\overline{P}$.

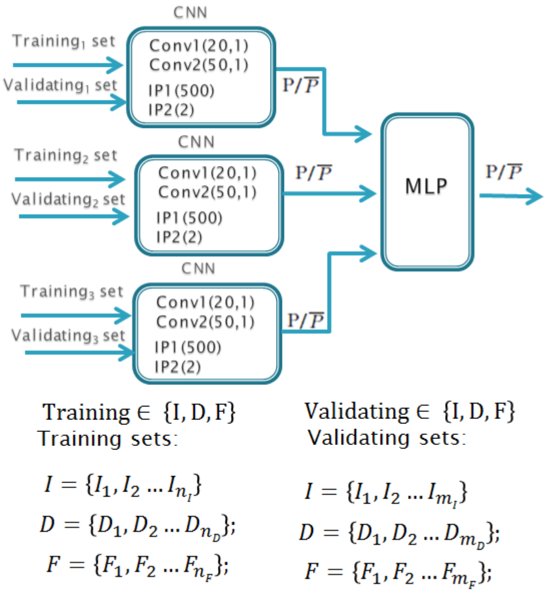

Figure 1: Late Fusion of Intensity, Depth and Flow Modalities

4 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 and 41834 for testing. The performances are measured by the Accuracy (ACC). The best performances optimized on the validation set were acquired with 29760 epochs and 0.01 learning rate. We start by comparing for each modality images the classification performances with LeNet architecture with different learning algorithms and we concluded the best performance measured was achieved the RMSPROP algorithm learning, with polynomial decay rate policy [5]. In Table 1 we show the ACC obtained with classical-training versus cross-training. The best performance is obtained for intensity (ACC = 96.55%) followed by depth and flow.

Table 1: Performance with classical-training and cross-training on Daimler testing set

Trained on	Validation on	Tested on	ACC
Intensity	Intensity	Intensity	**96.55%**
Intensity	Depth	Intensity	96.31%
Intensity	Flow	Intensity	96.23%
Depth	Depth	Depth	89.1%
Depth	Intensity	Depth	89.00%
Depth	Flow	Depth	**89.33%**
Flow	Flow	Flow	85.69%
Flow	Intensity	Flow	86.12%
Flow	Depth	Flow	**86.64%**

5 Conclusions

In this paper, we proposed different cross training approaches to improve pedestrian recognition. The cross-training approach performs slightly better the classical-training approach, but only for Flow and Depth modality. We are currently working on the late fusion architecture with RMSPROP algorithm learning and polynomial decay rate policy. For the future work, we will be concerned with improving that model by using optimal settings for different training modality sets and also by extending the model to cross datasets training.

References

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] Dănuţ Ovidiu Pop, Alexandrina Rogozan, Fawzi Nashashibi, and Abdelaziz Bensrhair. Fusion of stereo vision for pedestrian recognition using convolutional neural networks. *Proceedings of the European Symposium on Artificial Neural Networks (ESANN)*, 2017.