ESANN 2017 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 26-28 April 2017, i6doc.com publ., ISBN 978-287587039-1.
Available from http://www.i6doc.com/en/.

Fusion of Stereo Vision for Pedestrian Recognition using Convolutional Neural Networks

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

1 - INRIA Paris - RITS Team
Paris - France

2 - INSA Rouen - LITIS Laboratory
Rouen - France

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

Abstract. Pedestrian detection is a highly debated issue in the scientific community due to its outstanding importance for a large number of applications, especially in the fields of automotive safety, robotics and surveillance. In spite of the widely varying methods developed in recent years, pedestrian detection is still an open challenge whose accuracy and robustness has to be improved. Therefore, in this paper, we focus on improving the classification component in the pedestrian detection task on the Daimler stereo vision data set by adopting two approaches: 1) by combining three image modalities (intensity, depth and flow) to feed a unique convolutional neural network (CNN) and 2) by fusing the results of three independent CNNs.

1 Introduction

Pedestrian detection is a challenging task of great importance in the domain of object recognition and computer vision. It is a key problem for surveillance, robotics applications and automotive safety [1] where an efficient Advanced Driver Assistance System (ADAS) for pedestrian detection is needed to reduce the number of accidents and fatal injures[1]. These systems usually have a sensor network to capture the road data, and a signal/image processing component to extract pertinent features which are then classified by a recognition component.

A study performed by ABI Research published in 2015 shows that Mercedes-Benz, Volvo and BMW dominate the market for car enhancing ADAS systems. As from 2013, BMW cars have been fitted with a Driver Assistance package for Pedestrian Warning, based on infrared night-vision and monocular vision cameras. Recently, the Mercedes system has combined stereo vision cameras with long, medium and short-range radars to monitor the area in front of the vehicle. In 2016, the Continental company proposed an Advanced Radar Sensor (standard for VW Tiguan) able to detect both objects and pedestrians, at a

[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

distance of up to 170 meters. The Nissan company is developing a system which detects the car's environment, including the road, other vehicles and pedestrians.

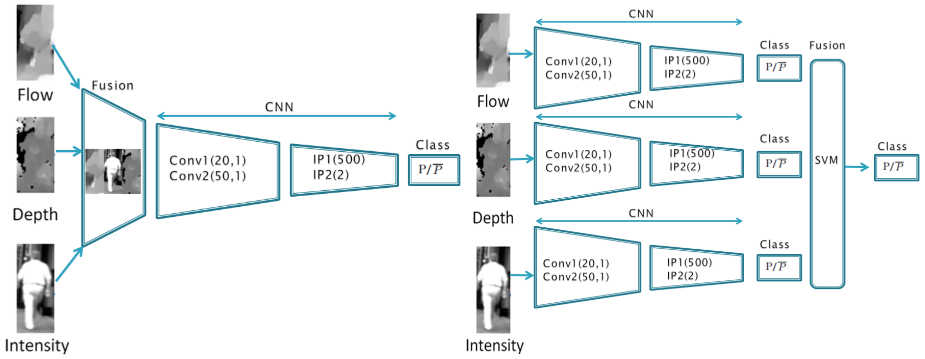
These existing ADAS systems still have difficulty distinguishing between human beings and nearby objects, especially in a crowded urban environment where they are not able to detect all partially occluded pedestrians, and they do not work efficiently in extreme weather conditions. Moreover, it is difficult to find an ADAS system that is able to ensure stable, real-time and effective full functionality.

In recent research studies, deep learning neural networks like LeNet, AlexNet, GoogLeNet have generally led to improvements in classification performance [2, 3, 4]. 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 level of risk.

Our work is concerned with improving the classification component of a pedestrian detector. The aim of the work presented in this paper is to train a CNN-based classifier using a combination of intensity, depth and flow modalities on the Daimler stereo vision data set. We compare two methods: early-fusion by combining image modalities to feed a unique CNN and respectively late-fusion of the results of three independent CNNs, one for each modality. We benchmark different learning algorithms and rate policies using the LeNet architecture. We show that the late-fusion classifier outperforms not only all single modalities but also the early-fusion classifier.

2 Previous work

Over the last decade, the pedestrian detection issue has been investigated intensely, resulting in the development of a widely varying detection methods were developed using a combination of features such as Integral Channel Features, Histograms of Oriented Gradients (HOG), Local Binary Patterns (LBP), Scale Invariant Feature Transform (SIFT), followed by a trainable classifier such as Support Vector Machine (SVM), Multilayer Perceptrons(MLP), boosted classifiers or random forests [5, 6]. In [7] the authors presented a mixture-of-experts framework performed with HOG, LBP features and MLP or linear SVM classifiers. Recently, in [8] a CNN to learn the features with an end-to-end approach was presented. This experiment focused on the detection of small scale pedestrians 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 [9]. A cascade Aggregated Channel Features detector is used in [10] to generate candidate pedestrian windows followed by a CNN-based classifier for verification 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 method combines the information of these modalities at the pixel level (early fusion), while the second architecture used separate subnetworks to generate a feature representation for each modality be-



(a) Early-Fusion Architecture (b) Late-Fusion Architecture

Fig. 1: Stereo Vision Modality Fusion Architectures. Conv(f,s) represents a convolution layer with f filters and a stride of s.

fore classification (intermediate fusion).

In this paper, we propose the fusion of the intensity, depth and flow modalities within a hybrid CNN-SVM framework. We believe that the late-fusion we propose based on three independent CNNs followed by an SVM is a promising approach and is more trainable (a sequential learning usually provides better results) than a huge complex architecture. For instance in [9] the authors used a full connection layer to make an intermediate-fusion, whereas we used an SVM classifier to fuse the CNN modality outputs.

3 The proposed architectures

We propose fusing stereo-vision information between three modalities: intensity, depth and flow. We investigate both early- and late-fusion architectures. For the early-fusion, we concatenate the corresponding modality images for deep learning within a unique CNN (see Fig. 1a). In the late-fusion, we train an SVM to discriminate between pedestrians (P) and non-pedestrians ($\overline{P}$) on the classification results of the three independent CNNs (see Fig. 1b).

We use 20 filters with one stride for the first convolution layer followed by 50 filters with one stride for the second. We use two Inner Product (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 $\overline{P}$.

We start by comparing AlexNet to LeNet with different learning algorithms: Stochastic Gradient Descent (SGD), Adaptive Gradient (ADAGRAD), Nesterov Accelerated Gradient (NAG), RMSPROP, ADAM and learning rate polices: Step Down (STEP), Polynomial Decay (POLY) and Inverse Decay (INV) on the intensity modality (see Table1). From the Caltech image data set we selected pedestrians bounding boxes (BB) of more than 50 px. All the BB were resized

AUC%	STEP		INV		POLY	
SGD	LeNet	95.69%	LeNet	96.24%	LeNet	95.6%
ADAGRAD	LeNet	94.17%	LeNet	92.66%	LeNet	94.64%
ADAM	AlexNet	92.24%	AlexNet	96.%	AlexNet	92.14%
RMSPROP	AlexNet	96.05%	AlexNet	93.05%	LeNet	**97.09%**
NAG	LeNet	96.98%	LeNet	95.82%	LeNet	95.51%

Table 1: Comparison of learning algorithms and rate policies on Caltech data set

	Intenity	Depth	Flow	Early-Fusion	Late-Fusion
AUC	96.39%	87.08%	88.24%	89.88%	99.21%
IC/2	±0.08	±0.15	±0.16	±0.14	±0.04

Table 2: Single-modality vs. multi-modality on Daimler testing set

to quadratic size (64 x 64 px). The best performance measured by Area Under the Curve (AUC) (97.09%) was achieved with the LeNet architecture using the RMSPROP learning[2], with POLY rate policy. We validate this benchmark result on the Daimler set where the best performance was also obtained with RMSPROP-POLY settings. We note that, with more complex problems, the breadth and depth of the network increases and becomes limited by computing resources, which thus hinders performance. [12, 13].

4 Experiments and Results

The training and testing were carried out on Daimler stereo vision images of 36 x 84 px with a 6-pixel border around the pedestrian image crops in three modalities: intensity, depth and optical flow, to allow fair comparisons [14].

We use 84577 samples for training, 75% of which were for learning, 25% for validation and 41834 for testing. The best performances optimized on the validation set were acquired with 208320 epochs and 0.01 learning rate. In Table 2 we show the AUC obtained with single-modality versus multi-modality. The best performance with single modality is obtained for intensity (96.39%) followed by depth and flow. For the multi-modality architectures, the late-fusion solution we propose not only outperforms all the single modality classifiers but also the early fusion solution. This improvement is statistically significantly since the confidence intervals are disjoint. These performance are also shown in the ROC curves (see Fig 2).

The classifier system proposed in [14] obtains an FPR of 0.0125% at 90% detection rate with their SVM early-fusion model on Daimler data set. Therefore our late-fusion model significantly improves the classification performance with Δ FPR=0.0085%. However the best classifier on the Daimler data set carried out with HOG+LBP/MLP, Monolithic HOG classifier (intensity+depth+flow) [7], allowed an FPR of 0.00026% at 90% detection rate. This system used features,

[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

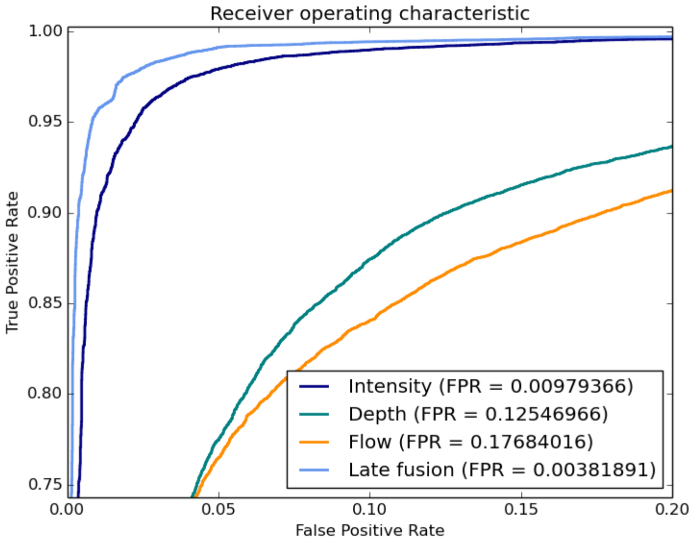


Fig. 2: Single-modality vs. multi-modality ROC classification performance on Daimler testing data set. FPR at 90% detection rate.

which are extracted explicitly and made the training more efficient while the CNN classifier has to extract implicit features and requires several samples for training.

5 Conclusion

In this paper, we focused on applying CNN models to improve the performance of the pedestrian recognizing component. This particular problem still remains an important challenge. We proposed the first application of LeNet architectures with RMSPROP algorithm learning for pedestrian recognition based on a multi-modal image data set. We also evaluated two CNN architectures, one for early- and the other for late-fusion. We showed that the late-fusion classifier outperforms the early-fusion one but does not go beyond the state-of-the-art. Experimental evaluations on the Daimler data set showed that the CNN fusion models represent promising approaches for multi-modality pedestrian recognition. The early fusion model has the advantage of having a less complex architecture with individual training on joint multi-model-images. The late fusion model achieves a better performance boost because it has a complex architecture which allows independent training of its components. For future work, we will improve our models with new CNN architectures, using transfer learning methods on different datasets.

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] 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.

[13] 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.

[14] 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.