ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8.
Available from http://www.i6doc.com/en/.

Multispectral Pedestrian Detection using Deep Fusion Convolutional Neural Networks

Jörg Wagner[1,2], Volker Fischer[1], Michael Herman[1] and Sven Behnke[2]

1- Robert Bosch GmbH - 70442 Stuttgart - Germany

2- University Bonn - Computer Science VI, Autonomous Intelligent Systems
Friedrich-Ebert-Allee 144, 53113 Bonn - Germany

Abstract. Robust vision-based pedestrian detection is a crucial feature of future autonomous systems. Thermal cameras provide an additional input channel that helps solving this task and deep convolutional networks are the currently leading approach for many pattern recognition problems, including object detection. In this paper, we explore the potential of deep models for multispectral pedestrian detection. We investigate two deep fusion architectures and analyze their performance on multispectral data. Our results show that a pre-trained late-fusion architecture significantly outperforms the current state-of-the-art ACF+T+THOG solution.

1 Introduction

Vision-based pedestrian detection is a crucial competence for many autonomous systems, such as self-driving cars or mobile robots. Although the topic has been intensively investigated in the last decade [1, 2, 3], it is still a challenging task. This is due to the variability of the environment as well as the variability between pedestrians, e.g. in shape or pose.

The majority of past research focused on the detection of pedestrians in visible spectrum images, where multiple benchmark datasets with comparatively large amounts of annotated pedestrians are available [1]. For a long period of time, approaches using hand-crafted features dominated these benchmarks. With the recent interest of the vision community in convolutional neural networks (CNNs), an increasing number of top performing detectors utilize CNNs.

A major drawback of visible image based pedestrian detectors is their poor performance at night time, as well as their sensitivity to illumination changes. To overcome these drawbacks, it is helpful to fuse the information of a visible camera with the information provided by a long-wavelength infrared (thermal) camera [3]. Due to the spectral band in which a thermal camera operates, it does not only omit the need for an external light source, but is also less affected by bad weather conditions. On the other hand, due to a high background temperature, thermal cameras often exhibit a decrease in image quality during daytime.

In the past, multispectral detectors (i.e. detectors which utilize the information of visible and thermal cameras) were mainly used in military and surveillance applications. With the recent decline in the price of thermal cameras, these detectors are becoming increasingly attractive for other applications.

Multispectral pedestrian detectors can be divided into three categories based on the level of abstraction at which the fusion takes place – pixel-level, feature-

ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8. Available from http://www.i6doc.com/en/.

level, and decision-level fusion. Choi et al. [4] apply a joint bilinear filter to fuse the thermal- and a visible image at pixel-level. In [5], a feature-level fusion based approach is developed by extending the aggregated channel features detector of [6]. Torresan et al. [7] detect and track pedestrians in the visible and thermal images separately and introduce an additional merging and validation process to combine the information at decision level.

This paper exploits deep model based detection methods, which have been successful in the visible domain, and extends these approaches to the multispectral case. To the best of our knowledge, it is the first attempt to fuse the images of a visible and thermal camera using deep models. We evaluate the introduced models and training methods on the KAIST multispectral pedestrain detection benchmark [5] and compare them to a state-of-the-art approach. We show that a late-fusion based deep model, which is additionally pre-trained, significantly outperforms the state-of-the-art baseline.

2 Multispectral Benchmark and Baseline

The KAIST multispectral pedestrian benchmark dataset [5] consists of temporally and spatially aligned visible and thermal images with a resolution of 640×512 pixels. In total, the dataset contains 95.3k pairs of visible-thermal images, split into a training set of 50.2k images with 41.5k labeled pedestrians and a test set of 45.1k images with 44.7k labeled pedestrians.

Currently, the best performing approach on the KAIST benchmark is an extension of the aggregated channel features (ACF) detector, which is introduced in [6]. The original ACF detector operates in a sliding window manner and uses subsampled and filtered channels as features. These channels are the components of the CIELUV color space, the normalized gradient magnitude and the histogram of oriented gradients. The multispectral extension of the ACF detector (ACF+T+THOG) additionally incorporates a contrast-enhanced version of the thermal images as well as HOG features of the thermal image as channels. The ACF+T+THOG detector which is used in our experiments is a retrained version of the current detector provided by the KAIST benchmark dataset. In the retrained version, we adjusted the standardization process of the ground truth bounding boxes. As suggested in [2], we standardize the bounding boxes by keeping the height and center fixed and only adjust the width.

3 Multispectral Deep Models

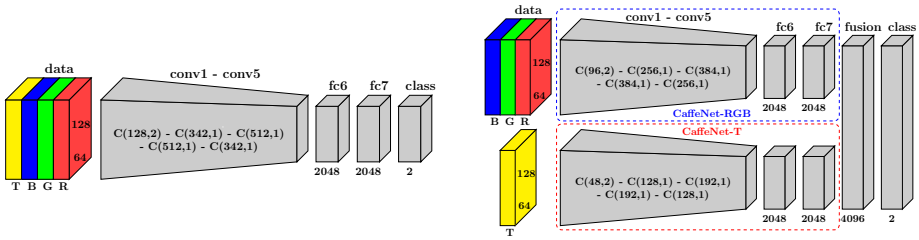
Our models are build upon the R-CNN detection framework [8], which is first applied in [9] to detect pedestrians. In the R-CNN framework, a proposal generator is used to generate candidate bounding boxes. Image data from these proposals is transformed into a standardized size and evaluated using a CNN.

To generate the proposals, we use the ACF+T+THOG detector. Based on these proposals, a CNN is used to fuse the information of the different modalities and to perform a binary classification.

ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8.
Available from http://www.i6doc.com/en/.

3.1 Fusion Architectures

We investigate an early- and a late-fusion based CNN architecture to fuse the information of the visible and thermal image (Figure 1). The early-fusion architecture (EarlyFusion) combines the information of the two modalities on pixel-level. In contrast, the late-fusion CNN (LateFusion) uses separate subnetworks to generate a feature representation for each modality. These representations are combined in an additional fully connected layer. Similar late-fusion architectures have been used in [10] to perform RGB-D object recognition and in [11] for action recognition in videos. The structure of the early-fusion architecture, as well as the subnetworks of the late-fusion architecture, are based on CaffeNet [12]. The following introduced network parameters are the result of a rough hand-tuned hyperparameter evaluation.



(a) Early Fusion (b) Late Fusion

Fig. 1: Architectures with changes compared to CaffeNet [12] visualized. *C(f,s)* represents a convolutional layer with f filters and a stride of s.

Early Fusion Architecture. For this architecture, we use CaffeNet and increase the number of filters per convolutional layer by a factor of $4/3$, to compensate the fourth input channel. Additionally, we decrease the number of neurons in the fully connected layers to 2048 and replace the 1000 class classification layer by a binary classification layer. Furthermore, the stride of the first convolutional layer is reduced to two, to obtain a sufficient spatial resolution after the last convolutional layer. By combining the two modalities at pixel-level, we expect a better utilization of inherent relations between the sensor modalities.

Late Fusion Architecture. The late-fusion architecture processes the data of the two modalities separately in subnetworks and fuses the resulting feature representations in a fully connected layer. Both subnetworks are based on CaffeNet without its classification layer and use, analogous to the early-fusion CNN, 2048 neurons in their fully connected layers, as well as a stride of 2 in their first convolutional layer. In the subnetwork, which processes the thermal images we additionally halve the number of filters per convolutional layer. The factor 0.5 is derived based on the number of grayscale filters in the first convolutional layer of CaffeNet. The resulting activations, generated by the second fully connected layer, of the two subnetworks are concatenated and fused in a fully connected layer with 4096 neurons. The fusion layer is followed by a ReLU non-linear layer, a dropout layer, and finally a binary classification layer. The parameters of the late-fusion network are learned in an end-to-end fashion.

ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8.
Available from http://www.i6doc.com/en/.

3.2 Training Procedure

The availability of a sufficiently large amount of labeled data is often a crucial point when training deep networks. Due to the cost associated with data generation and labeling, the amount of available training data is limited in most applications. One popular approach to overcome this problem is to pre-train the network on a large auxiliary dataset. To show the benefit of pre-training, we train our networks twice. Both by only using the training data provided by the multispectral benchmark and by additionally pre-training the networks on auxiliary datasets. Due to the lack of available large visible-thermal image datasets, we pre-train our models on visible data. The red channel is used as a crude approximation of the thermal channel. Thus, during pre-training, the images of the early-fusion architecture contain the red channel twice. This approximation may be too rough, given the difference between red and thermal images in real data. Furthermore, it is questionable whether the early-fusion architecture can learn complementary features which utilize the substituted thermal channel.

Our pre-training procedure consists of the following two steps: In the first step, the convolutional layers of the CaffeNet-T, CaffeNet-RGB and EarlyFusion network are trained on the task of image classification using the ImageNet [13] dataset. In the second step, we fine-tune these networks using all images of the CALTECH benchmark [2]. Therefor we adopt the weights of the already trained convolutional layers and initialize the weights of the fully connected layers with random values. Additionally, we add a randomly initialized classification layer to each of these networks. During the aforementioned pre-training steps, the two subnetworks of the late-fusion architecture are solely trained separately.

The training of the late-fusion model on the KAIST data occurs in two steps as well: Depending on whether we use pre-training or not, the two subnetworks of the LateFusion architecture are initialized with either pre-trained weights, or random values. Starting from these parameters, we separately optimize the two subnetworks. The second step encompasses a joint fine-tuning of the whole late-fusion architecture. As suggested by [10], the best fusion results can be reached when the weights of the subnetworks are fixed and only the fusion layers are trained. Due to its simple structure, the early-fusion network can be trained or fine-tuned, depending on the usage of pre-training, in a regular fashion.

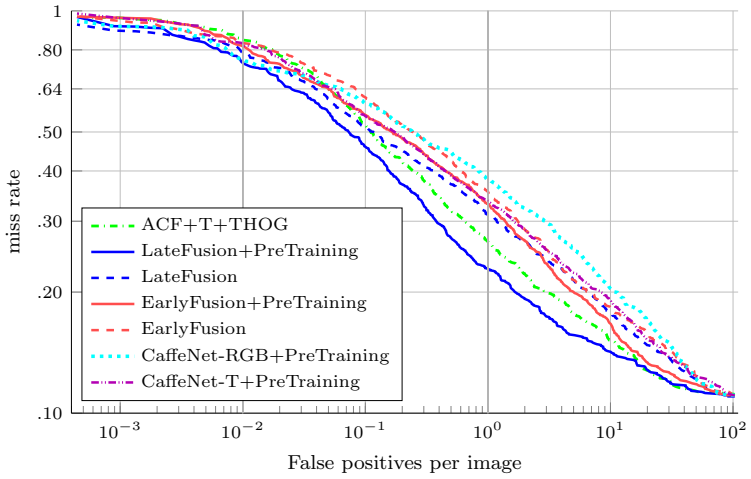
In contrast to the data sampling policy of the baseline detector, we analogously to [9] use every second frame of the KAIST training dataset. Additionally, we split the original training data into a training set containing 92 % of the images and a validation set containing the remaining 8 % of the images.

4 Results

The evaluation of the detectors is performed on the *reasonable* subsets of the KAIST test data. The *reasonable day* and *reasonable night* subset respectively contain images captured during daytime and nighttime, the *reasonable all* data is formed by the union of these to datasets. We follow the evaluation protocol defined in [5] and additionally standardize in accordance to [2] the aspect ra-

ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8. Available from http://www.i6doc.com/en/.

tio of bounding boxes to a fixed value of 0.41, by means of width adjustment. Figure 2 shows the ROC curves of the detectors, as well as their log-average miss rates, as defined in [2]. For each fusion architecture, we report the performance with and without pre-training, marked by the PreTraining suffix. In addition, the performance of the two pre-trained late-fusion subnetworks CaffeNet-T+PreTraining and CaffeNet-RGB+PreTraining is visualized. The pre-trained late-fusion based deep architecture significantly outperforms the state-of-the-art ACF+T+THOG baseline, as well as all other evaluated detectors in all three test set splits. At daytime, the performance of the LateFusion+PreTraining architecture is 5.12 % better than the baseline and at nighttime 10.4 %. Even though during pre-training the substitution of the thermal channel with the red channel is a very rough approximation, it leads to a significant performance increase in all architectures. Most of the time, the early-fusion architecture is not able to reach state-of-the-art ACF+T+THOG performance. In our opinion, there are at least three reasons for this observation: The poor performance of the early-fusion network without pre-training can most likely be traced back to the limited amount of data available in the KAIST dataset. Additionally, we suspect that the early-fusion network did not learn meaningful multimodal features during the pre-training process, due to the absence of complementary information in



(a) Detector ROC curves on the *reasonable all* data split

Detector	log-average miss rate		
	all	*day*	*night*
ACF+T+THOG	50.48 %	51.27 %	47.40 %
LateFusion+PreTraining	**43.80 %**	**46.15 %**	**37.00 %**
LateFusion	51.30 %	55.27 %	41.58 %
EarlyFusion+PreTraining	53.94 %	50.90 %	51.76 %
EarlyFusion	57.96 %	58.22 %	57.78 %
CaffeNet-RGB+PreTraining	56.52 %	53.51 %	63.40 %
CaffeNet-T+PreTraining	54.67 %	59.77 %	42.09 %

(b) Log-average miss rate on the *reasonable all*, *day* and *night* data split

Fig. 2: Detection performance comparison of the introduced fusion architectures, with the recent state-of-the-art detector ACF+T+THOG.

ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8. Available from http://www.i6doc.com/en/.

the chosen approximation of the thermal channel. A third reason for the poor performance of the early-fusion architectures is a small non-systematic alignment error, which is noticeable in some images of the dataset. The late-fusion network is able to cope with such errors, because it fuses information at a stage where spatial information is less relevant. As expected, the thermal modality has advantages over the visible at nighttime and vice versa at daytime.

5 Conclusion

We introduced a first application of deep CNNs for pedestrian detection on the basis of multispectral image data and evaluated two deep architectures, one for early- and the other for late-fusion. Our analysis on the KAIST Multispectral Benchmark dataset shows that a pre-trained late-fusion based architecture can significantly outperform the state-of-the-art ACF+T+THOG solution, whereas most of the time the early-fusion architecture is not able to reach state-of-the-art performance. This may be due to the inability of the early-fusion network to learn meaningful multimodal abstract features in the given setting.

References

[1] R. Benenson, M. Omran, J. Hosang, and B. Schiele. Ten years of pedestrian detection, what have we learned? In *CVRSUAD, ECCV Workshop*, 2014.

[2] P. Dollár, C. Wojek, S. Bernt, and P. Perona. Pedestrian detection: An evaluation of the state of the art. *IEEE Trans. on Pattern Anal. and Mach. Intell.*, 34(4):743–761, 2012.

[3] R. Gade and T. B. Moeslund. Thermal cameras and applications: A survey. *Machine Vision Applications*, 25(1):245–262, 2014.

[4] E.-J. Choi and D.-J. Park. Human detection using image fusion of thermal and visible image with new joint bilateral filter. In *ICCIT*, pages 882–885, 2010.

[5] S. Hwang, J. Park, N. Kim, Y. Choi, and I.S. Kweon. Multispectral pedestrian detection: Benchmark dataset and baseline. In *CVPR*, 2015.

[6] P. Dollár, R. Appel, S. Belongie, and P. Perona. Fast feature pyramids for object detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 36(8):1532–1545, 2014.

[7] H. Torresan, B. Turgeon, C. Ibarra-castanedo, P. Hébert, and X. Maldague. Advanced surveillance systems: Combining video and thermal imagery for pedestrian detection. In *In Proc. of SPIE, Thermosense XXVI, volume 5405 of SPIE*, pages 506–515, 2004.

[8] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014.

[9] J. Hosang, R. Benenson, M. Omran, and B. Schiele. Taking a deeper look at pedestrians. In *CVPR*, 2015.

[10] A. Eitel, J. T. Springenberg, L. Spinello, M. Riedmiller, and W. Burgard. Multimodal deep learning for robust rgb-d object recognition. In *IROS*, 2015.

[11] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *NIPS*, pages 568–576, 2014.

[12] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

[13] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *IJCV*, 115(3):211–252, 2015.