Pedestrian Detection based on Deep Fusion Network using Feature Correlation

Yongwoo Lee, Toan Duc Bui and Jitae Shin
School of Electronic and Electrical Engineering, Sungkyunkwan University, Suwon, South Korea
E-mail: {tencio2001, toanhoi, jtshin}@skku.edu Tel: +82-31-290-7994

***Abstract*— Since most of the pedestrian detection method focus on color images, the detection accuracy is lower when the images are captured at night or dark. In this paper, we propose a deep fusion network based pedestrian detection method. We utilize deconvolutional single shot multi-box detector (DSSD) fused at halfway stage. Also, we apply feature correlation for two image modality feature maps to produce a new feature map. For the experiment, we use KAIST dataset to train and test the proposed method. The experiment results show that the proposed method gains 22.46% lower miss rate compared to the KAIST pedestrian detection baseline. In addition, the proposed method shows at least 4.28% lower miss rate compared to the conventional halfway fusion method.**

I. INTRODUCTION

Object detection is one of the most fundamental research area in computer vision. With the success of deep learning, many object detection research deploy deep learning based method to improve the detection accuracy. As a sub-category in the object detection research, pedestrian detection is very attractive research. Especially nowadays, autonomous driving is one of the major research in computer vision and pedestrian detection is essential to avoid possible accidents. It is a fact that deep learning technique improves the detection accuracy, however, the gap between perception of humans and machines is still big [1]. The accuracy becomes worse when the object is occluded, the resolution is low, and the background is complex. Moreover, most of the current research are focused on the color image based detection. It often shows lower accuracy when the test sequences is captured at night or the object goes into the shade. It is the main weakness of vision-based object detection since autonomous vehicles should be able to detect the object when it is day or night.

One can solve the issue by using color images with multiple sensors such as depth map camera or thermal camera. The images captured by thermal camera have less influence by the light. Therefore, color images and thermal images have complementary effect if they are used together. However, as we mentioned earlier, most of the research is focused on the single modality data, which is often color image. Recently, a few research proposed multispectral object detection method using deep learning technique. J. Wagner et al. proposed region-based convolutional neural networks (R-CNN) [2] based fusion networks [3]. In early fusion, color and thermal image is fused to have four channels to be fed into R-CNN. In late fusion, color and thermal images are fed into separate R-CNN and the output of the last fully-connected layer is fused to classify the object. However, R-CNN has lower performance compared to recent techniques [4-6] and the fusion network architecture is too simple. J. Liu et al. employed Faster R-CNN [5], which is the next version of R-CNN, to propose the fusion networks [7]. In this paper, four fusion networks are introduced and investigated which position is the best location to fuse the networks. Faster R-CNN significantly reduced prediction time compared to R-CNN. However, the improvement for the prediction accuracy is relatively small.

Meanwhile, there are some attempts to reduce the prediction time while maintaining the accuracy over a certain point. J. Redmon et al. proposed You Only Look Once (YOLO) to solve the issue with prediction time [8]. Recently, they proposed YOLO version 3 (YOLOv3) to increase the accuracy and yet still fast enough [9]. After YOLO first version was proposed, W. Liu et al. proposed Single Shot Multi-box Detector (SSD) [10]. They also focused on fast prediction rather than increasing the prediction accuracy thoroughly. The next version, deconvolutional SSD (DSSD) was proposed in the last year to improve the prediction accuracy [11]. These method are capable of real-time prediction since they do not have independent networks to predict the bounding box for the objects. In R-CNN based techniques, the region proposal networks (RPN) predicts possible bounding boxes for the objects. Around 2,000 candidates are used, which slow down the whole prediction procedure. However, YOLO and SSD do not have the networks as RPN. For this reason, they are suitable method to apply the real-time system as autonomous driving.

Therefore, in this paper, we propose a method that takes advantage of the fusion networks and fast prediction object detection method with feature correlation. For the fusion network, we apply halfway fusion proposed in [7]. For the deep learning network, we apply DSSD which is the second version of SSD. We selected DSSD over YOLOv3 for the following reasons. First, DSSD is built upon Caffe [12], YOLOv3 is built upon Darknet [13]. Caffe and Darknet are deep learning frameworks to develop machine learning application. Caffe is very popular deep learning framework that supports at most features compared with the other frameworks. Darknet, on the other hand, is proposed with YOLO and only a few research groups use the framework. Therefore, it is not easy to tryout the new network architecture in the Darknet. Secondly, YOLOv3 is rather focused on object classification method. There second version is YOLO 9000 [14] since it can classify over 9000 class

with their improved classification method. However in our application, we only detect pedestrian so that the number of class is only two. Therefore, the benefit from YOLOv3 is not greatly applied to our case.

The rest of the paper is organized as follows. In section 2, we introduce the fusion networks proposed in [3, 7] and DSSD network architecture. In section 3, we propose the fusion network with DSSD networks and feature correlation. We show the experimental results in section 4, and finally, we conclude the paper in section 5.

II. Related Works

In this section, we investigate the fusion networks for color and thermal images in [3, 7]. Also we introduce DSSD network architecture.

A. Fusion Networks

J. Wagner et al. proposed R-CNN based fusion networks [3]. In this early study, the networks are fused in the beginning and in the end of VGG-16 networks. VGG-16 has five convolution layer blocks (Conv) and two fully connected layers (FC) [16]. In the early fusion, three channels of color image and one channel of thermal image are concatenated to have four channels and fed into R-CNN network. In the late fusion, each modality is fed into independent R-CNN network. After the last fully connected layer, both outputs are accumulated and they are used to decide to class information.

J. Liu further investigated the location for the fusion networks [7]. In the paper, they proposed four fusion networks, which are early, halfway, late, and score fusion. Fig. 1 depicts the four network architectures. In early fusion, the two modalities are fed into different convolution layer. After the first Conv in Faster R-CNN, two feature maps are concatenated to be fed into the rest of the convolution layers. The concatenated feature map has twice channel than the original one. To reduce the channel, Network-in-Network (NIN) layer is applied to make the channel size as half [17]. In this case, one can use pre-trained coefficient. In halfway fusion, the location for the feature concatenation is after the fourth Conv, in late fusion, it is after the last FC. Score fusion is a cascade of two Faster R-CNN. The output from the color detection, is sent into the thermal networks to obtain detection score, and vice versa. In early fusion, the feature maps are concatenated after the first Conv, lower level features are concatenated. In halfway fusion, since feature maps after fourth Conv are concatenated, they have more semantic information as well as fine detail. In late fusion, the output from the last FC is concatenated and it is also higher level features. In [7], halfway fusion shows the best results among the four method.

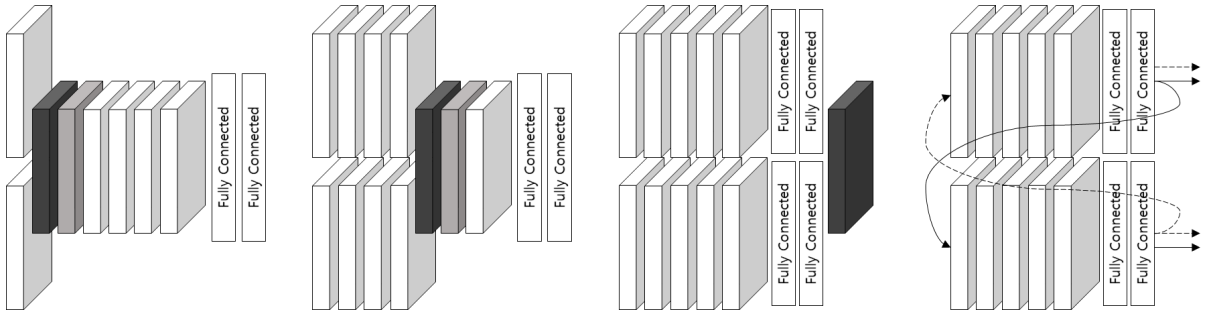
B. DSSD

DSSD is the second version of SSD. The main feature of the work is to detect the object very fast. The main difference between R-CNN based techniques and SSD is that SSD do not have RPN for the bounding box prediction. Since RPN takes much time in the whole prediction process, they resolutely remove RPN. Instead, they take the use of default boxes which have pre-defined aspect ratio. Also, their default boxes have different scales to decrease the localization error. Fig. 2 shows the network architecture of SSD. They replaced FC6 and FC7 with Conv and added extra feature layers.

Based on their success on SSD, DSSD replace VGG-16 with Residual-101 [18] and add more features to improve the detection accuracy. Since SSD tends to have lower accuracy for small objects, they utilize deconvolution layers in the next version. If one use deconvolution layers to have symmetric sandglass structure, the prediction time will increase. Therefore, they apply asymmetric architecture because their first concern is fast prediction. Fig. 2 shows the extra DSSD layers. The base network is replaced with Residual-101 in DSSD.

III. Proposed Method

In this section, we propose deep fusion network using color and thermal images to solve issue with color image based detection method.



(a) Early fusion b) halfway fusion (c) late fusion (d) score fusion

Fig. 1 the network fusion architectures in [7]: white, gray and black boxes indicates convolution layer, NIN layer, and concatenate layer, respectively.

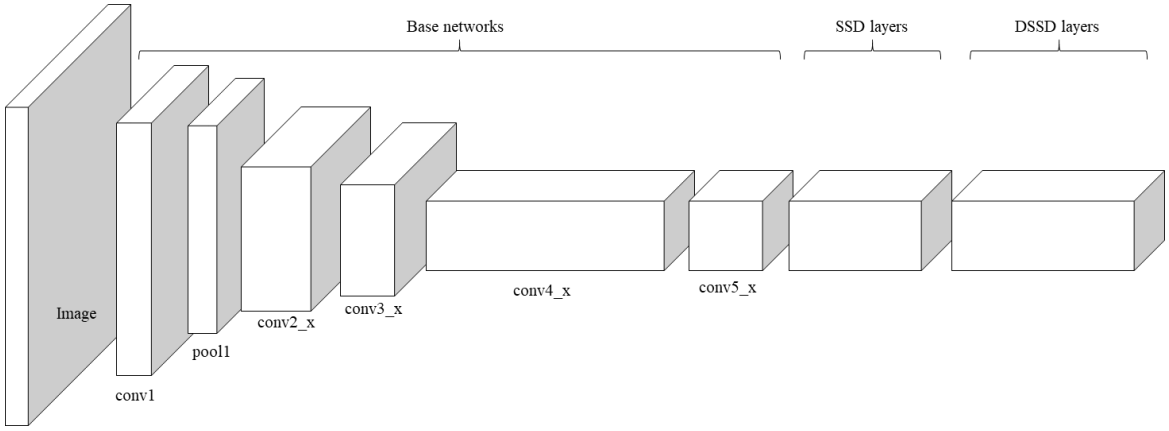


Fig. 2 DSSD network structure: Base networks are Residual-101 for DSSD and VGG-16 for SSD. SSD layers include 5 convolution layers with different scale. DSSD layers include 5 deconvolution layers. SSD do not have DSSD layers, DSSD have both layers.

A. Fusion Networks with DSSD

We use DSSD, which is SSD version 2, as a main network architecture for the fusion networks. We adapt halfway fusion since halfway fusion is the best among the four fusion method. In Fig. 3, we show the proposed network architecture. The expression Conv*N*_*M* in the figure means *M*-th convolution layers in *N*-th convolution block. In DSSD, VGG-16 is replaced with Residual-101. Therefore we have to choose the proper location for the fusion network. Adapted from [7], Conv4_22 is selected for the location to concatenate the feature maps.

B. Feature Correlation Layer

To improve detection accuracy, we apply feature correlation layer. One can see that the color images and thermal images have considerable correlation. Also, we are interested in pedestrian detection that usually have higher pixel values in the thermal image. If the color and thermal images are well calibrated, the region with higher intensity in the thermal image is emphasized in the output feature map. Therefore it is helpful to improve the detection accuracy even more. Given two feature maps h_{color} and h_{thermal}, the output of the feature correlation layer is defined as:

$$h_{\text{corr}} = \sqrt{h_{color} \circ h_{thermal}} \tag{1}$$

where $\circ$ is the Hadamard product [22]. It is element-wise product between the feature maps from the different modality network stream. After the correlation feature map is produced, it is sent to the concatenate layer. The inputs for the concatenate layer are feature maps from each modality network stream and the correlation feature map. NIN layer reduce the number of channels to use the pre-trained coefficients. The rest of the network is the same as original.

IV. Experimental Results

We used KAIST multispectral pedestrian dataset for the experiment [15]. The dataset has totally 95,328 color and thermal image pairs and 1,182 pedestrians are appeared with 103,128 annotations. In the training part, we excluded the images if the pedestrians are occluded, or if it is less than 50 pixels. Every two frames are extracted, totally 3,357 pairs are used for the training. In the testing part, we applied the same condition and extracted every three frames. Totally 2,094 pairs are used to evaluate the method. For the training parameters, we applied same settings as in [11] mostly. We changed the iteration number as 150K since the dataset has different characteristics compared to PASCAL VOC dataset. Also, we changed the learning rate criterion. The starting learning rate is 10^{-3} and it is decreased to 10^{-4} at 100K and 10^{-5} at 130K. The trained model for SSD layer is then used as pre-trained model for DSSD.

To evaluate the results, we used FPPI-miss rate metric. ACF+T+THOG is KAIST pedestrian detection baseline method [15]. It is an object detection method which has 10 channels of conventional ACF feature [19] plus features from thermal image and HOG features [20] from the thermal images. Halfway fusion uses color and thermal images applied to faster R-CNN with VGG-16. We also compared with SSD networks fused in halfway (SSD-H). In Fig. 4, we show FPPI-miss rate results to compare the four method. It is compared in $[10^{-2}, 10^{0}]$ log scale range for the miss rate, the average is marked at 10^{-1} as proposed in [21]. The lower is more accurate.

The proposed DSSD halfway with feature correlation (DSSD-HC) shows better performance compared to the other method. It shows 22.46% better miss rate compared to ACF+T+THOG. Compared to the conventional halfway fusion, our method shows 6.38% better performance. In our previous work, using SSD shows 38.60% of miss rate. Our newly proposed method shows 34.32% of miss rate which is far better than the previous work. It is because DSSD increased about 2-3% of detection accuracy and the feature correlation layer increased about 1-2%.

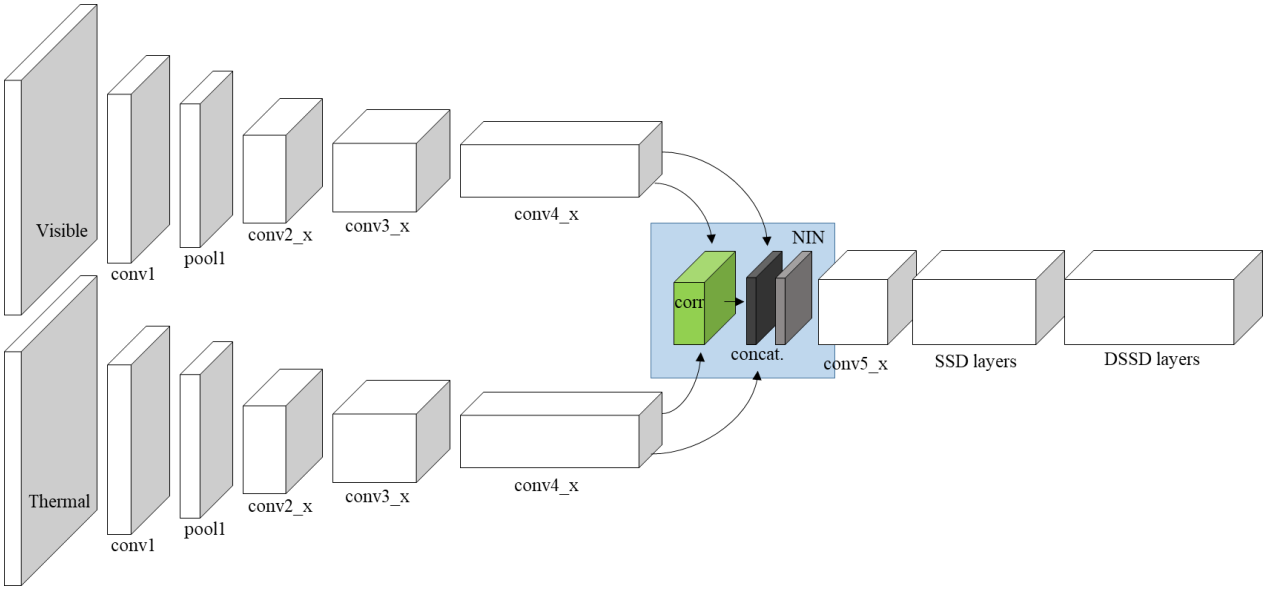


Fig. 3 The proposed fusion networks. Note that correlation and concatenate layers are located after Conv4_22. The proposed method is highlighted in blue box.

In this paper, we chose halfway fusion because halfway fusion is middle level fusion, and the fusion features have semantic information as well as fine details. Early fusion concatenate the feature maps close to the pixel level so that it cannot have enough benefit of fusion networks. Late fusion concatenate the semantic feature maps. If the semantic feature maps have inaccurate information for the classification, the total performance is affected seriously by the wrong information.

For the perception evaluation, we compared the proposed results with DSSD detection results using color images only for the training and testing. In this results, one can see the benefit of fusion networks when it comes to the images captured at night or the images have pedestrian appeared in the dark region. In Fig. 5, the first two rows show DSSD detection results using color images only and the last two rows shows the proposed method. In the figure we can see that the proposed method detects better compared to the single modality method. Even in the images with good lighting condition, the detection results are better since we can utilize the extra information from the thermal images. For the images captured at night or dark, thermal images give solid information about the pedestrian. Therefore, even if the color image cannot produce good feature maps, the proposed method can detect the pedestrian accurately.

V. CONCLUSIONS

In this paper, we proposed a deep-fusion network based pedestrian detection method using color and thermal images with feature correlation. The conventional method only utilize color images so that detection accuracy become lower when the images are captured at night or the objects are in the dark region. Instead, our proposed method deploy color and thermal image pairs to solve the issue with lighting condition. Also, we proposed correlation feature layer to amplify the benefit of using two modalities. Experimental results show the proposed method show 22.46% better FPPI-miss rate compared to the KAIST pedestrian detection baseline method. In addition, the proposed method show at least 4.28% better FPPI-miss rate compared to the conventional halfway fusion method.

For the future works, we will investigate the right location for the correlation layer in the network. To improve detection accuracy, we will focus on the benefit of using two modalities and suggest new fusion network method.

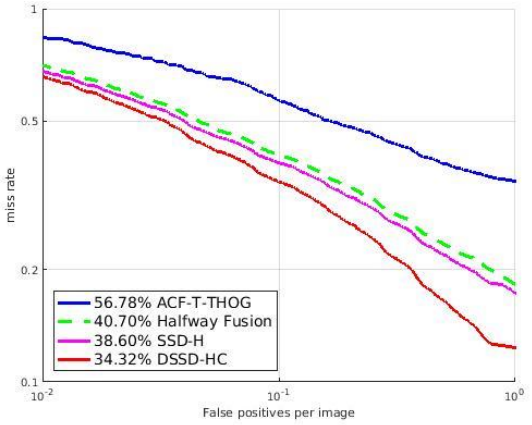


Fig. 4 FPPI-miss rate results comparison

(a) (b) (c) (d) (e)

Fig. 5 The perception evaluation to compare the DSSD trained with color images (first two rows) and DSSD-HC (bottom two rows). DSSD-HC accurately detects the pedestrian in all images. DSSD trained with color images detects wrong object especially when the images are dark.

ACKNOWLEDGMENT

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2017R1D1A1B03031752, NRF-2018R1C1B6007462).

This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(2018-0-01798) supervised by the IITP(Institute for Information & communications Technology Promotion)

REFERENCES

[1] S. Zhang, R. Benenson, M. Omran, J. Hosang, and B. Schiele, "How far are we from solving pedestrian detection?," *IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, USA, pp.1259-1267, 2016.

[2] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," *IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, USA, pp. 580-587, 2014.

[3] J. Wagner, V. Fischer, M. Herman, and S. Behnke, "Multispectral pedestrian detection using deep fusion convolutional neural networks," *European Symposium on Artificial Neural Networks*, Bruges, Belgium, pp. 509-514, 2016.

[4] R. Girshick, "Fast r-cnn," *arXiv preprint arXive:1504.08083*, 2015.

[5] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *Neural Information Processing Systems, Montreal*, Canada, pp. 91-99, 2015.

[6] K. He, G. Gkioxari, P. Dollar, R. Girshick, "Mask R-CNN," *IEEE International Conference on Computer Vision*, Venice, Italy, pp. 2980-2988, 2017.

[7] J. Liu, S. Zhang, S. Wang, D. N. Metaxas, "Multispectral deep neural networks for pedestrian detection," *arXiv preprint arXiv:1611.02644*, 2016.

[8] Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 779-788).

[9] Redmon, J., & Farhadi, A. (2018). YOLOv3: An incremental improvement. *arXiv preprint arXiv:1804.02767*.

[10] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. –Y. Fu, and A. C. Berg, "SSD: Single Shot MultiBox Detector,"

European Conference on Computer Vision, Amsterdam, the Netherlands, pp. 21-37, 2016.

[11] Fu, C. Y., Liu, W., Ranga, A., Tyagi, A., & Berg, A. C. (2017). DSSD: Deconvolutional single shot detector. *arXiv preprint arXiv:1701.06659*.

[12] Jia, Yangqing, et al. "Caffe: Convolutional architecture for fast feature embedding." *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 2014.

[13] Redmon, Joseph. "Darknet: Open source neural networks in c." *Pjreddie. com.[Online]. Available: https://pjreddie. com/darknet/.[Accessed: 21-Jun-2017]* (2016).

[14] Redmon, Joseph, and Ali Farhadi. "YOLO9000: better, faster, stronger." *arXiv preprint* (2017).

[15] S. Hwang, J. Park, N. Kim, Y. Choi, and I. S. Kweon, "Multispectral pedestrian detection: benchmark dataset and baseline," *IEEE Conference on Computer Vision and Pattern Recognition*, Boston, USA, pp. 1037-1045, 2015.

[16] K. Simonyan, A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[17] M. Lin, Q. Chen, S. Yan, "Network in Network," *arXiv preprint arXive:1312.4400*, 2013.

[18] He, Kaiming, et al. "Deep residual learning for image recognition." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.

[19] P. Dollar, R. Appel, S. Belongie, and P. Perona, "Fast feature pyramids for object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 36, No. 8, pp. 1532-1545, Jan. 2014.

[20] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," *IEEE Conference on Computer Vision and Pattern Recognition*, San Diego, USA, pp. 886-893, 2005.

[21] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: A benchmark," *IEEE Conference on Computer Vision and Pattern Recognition*, Miami, USA, pp. 304-311, 2009.

[22] Horn, Roger A. "The hadamard product." *Proc. Symp. Appl. Math*. Vol. 40. 1990.