IET Computer Vision

Research Article

IET Journals
The Institution of Engineering and Technology

Multi-layer fusion techniques using a CNN for multispectral pedestrian detection

ISSN 1751-9632
Received on 5th January 2018
Revised 18th May 2018
Accepted on 05th July 2018
E-First on 15th August 2018
doi: 10.1049/iet-cvi.2018.5315
www.ietdl.org

Yunfan Chen[1], Han Xie[1], Hyunchul Shin[1] ✉

[1]*Department of Electronics and Communication Engineering, Hanyang University, Ansan, Republic of Korea*
✉ *E-mail: shin@hanyang.ac.kr*

Abstract: In this study, a novel multi-layer fused convolution neural network (MLF-CNN) is proposed for detecting pedestrians under adverse illumination conditions. Currently, most existing pedestrian detectors are very likely to be stuck under adverse illumination circumstances such as shadows, overexposure, or nighttime. To detect pedestrians under such conditions, the authors apply deep learning for effective fusion of the visible and thermal information in multispectral images. The MLF-CNN consists of a proposal generation stage and a detection stage. In the first stage, they design an MLF region proposal network and propose to use summation fusion method for integration of the two convolutional layers. This combination can detect pedestrians in different scales, even in adverse illumination. Furthermore, instead of extracting features from a single layer, they extract features from three feature maps and match the scale using the fused ROI pooling layers. This new multiple-layer fusion technique can significantly reduce the detection miss rate. Extensive evaluations of several challenging datasets well demonstrate that their approach achieves state-of-the-art performance. For example, their method performs 28.62% better than the baseline method and 11.35% better than the well-known faster R-CNN halfway fusion method in detection accuracy on KAIST multispectral pedestrian dataset.

1 Introduction

Pedestrian detection as a canonical sub-problem of object detection has received great attention during recent years, since it is an essential technique for various applications such as video surveillance [1], person identification [2], people tracking [3], and advanced driver assistance systems (ADASs). Despite extensive efforts for solving pedestrian detection problems, it is still regarded as a challenging problem. This is due to tiny and occluded appearances, cluttered backgrounds, and adverse illumination conditions. However, most past research have been limited to good lighting conditions. As a result, detecting pedestrians in the case of illumination variation, shadows, and low external light at nighttime is still a challenging problem.

To overcome these problems, it is helpful to fuse the information of a visible camera with the information provided by a thermal camera. A thermal camera is useful under adverse illumination conditions and is less reliant on the surrounding lighting changes. However, it loses fine visual details of human objects (e.g. clothing) that can be captured by visible cameras depending on external illumination. By implementing sensor fusion, these varying, but complementary characteristics can be integrated efficiently for robust environment perception. As shown in Fig. 1*a*, the visibility of pedestrians in visible images is limited due to poor illumination conditions. However, in thermal images, the intensity and shape information of pedestrians can be provided. As can be seen in Fig. 1*b*, it is difficult to distinguish colour and detailed information on pedestrians' clothes, which can be provided by visible images. In a bright background, the visible image provides more distinctive visual features for the pedestrians against background objects. In such a scenario, human silhouettes in a thermal image are ambiguous, as shown in Fig. 1*c*.

Recently, multispectral detectors (i.e. detectors that utilise the information of visible and thermal cameras) are becoming increasingly attractive for many applications such as military, ADAS, surveillance etc. Furthermore, great progress has been made by deep convolutional neural networks (CNNs) on pedestrian detection [4–10]. These advantages make it very natural and interesting to exploit the effectiveness of CNNs for multispectral pedestrian detection. Therefore, several deep NN models have been proposed to integrate information from multimodal data sources [11–13] such as image versus audio, image versus text, and image versus video. However, there have only been a few studies on how to apply CNNs to vision problems with multispectral data sources, except for very recent efforts [14–17]. Therefore, how visible and thermal image channels can be properly fused in CNNs to achieve the best performance in pedestrian detection remains to be solved.

In this paper, we focus on developing a CNN fusion architecture that is able to take full advantage of synergy with visible and thermal images for multispectral pedestrian detection. It is likely that these two streams of CNN-based detectors, if fused appropriately, could provide complementary detection information and hence there is a big potential to improve the detection performance by leveraging multispectral images. However, it is not easy to explore the most effective CNN architecture that simultaneously capitalise the information in both of visible and thermal images for pedestrian detection. This is because CNN consists of several layers, and features at different layers correspond to various levels of semantic meanings and fine visual details. Fusion at different layers with different fusion algorithms would lead to different detection results. Therefore, it is very important to find where to fuse and how to fuse the two networks to make the best use of the multispectral images. From our systematic experimental analysis, we propose a novel architecture for fusion of two streams of networks that achieve state-of-the-art performance. Our proposed architecture consists of a proposal network and a detection network. The main contributions of this work can be summarised as follows.

First, we developed a new fusion technique devising a new multiple-layer fusion architecture. Our architecture adopted a region proposal network (RPN), which is a type of CNN. This multi-layer fusion method outperforms the baseline method on multispectral pedestrian dataset [18]. This implies that the diverse characteristics of different layers, low-level features, visual details, and semantic features can provide richer feature abstractions, which is the best fusion strategy.

Second, we found that our proposed multi-layer fused (MLF) RPN with summation fusion provides the best performance on multispectral pedestrian detection. Previous methods used concatenation instead of summation.

Fig. 1 *Examples of thermal images (top) and visible images (bottom) for visualising complementary characteristics of multispectral data*

Third, based on MLF-RPN, we further improve the detection stage to generate our final MLF-CNN. We take features not from one separate layer such as previous works, but from multiple feature maps. Then, the region of interest (ROI) pooling layers of each feature map are combined for the later classification. Experimental results show that the extra ROI pooling fusion process can further improve the accuracy.

Finally, our proposed MLF-CNN achieves state-of-the-art performance on several challenging datasets. Our method achieves 11.35, 12.17, and 20.01% lesser miss rates on the KAIST, UTokyo multispectral datasets, and OSU colour–thermal dataset [18–20], respectively, compared with the well-known faster regions with CNN features (R-CNN) halfway fusion method [14].

The rest of this paper is organised as follows. In Section 2, related works are briefly reviewed. In Section 3, our proposed MLF-CNN is described in detail. In Section 4, experimental results and analysis are presented. Finally, conclusions and future works are summarised in Section 5.

2 Related works

We first give a review of pedestrian detection under good illumination conditions, and then discuss the related works of multispectral pedestrian detection under adverse illumination conditions.

Methods of pedestrian detection can be divided into two main categories: handcrafted channel-based methods and CNN-based methods. Handcrafted channel-based methods require manually designed features to express images. On the other hand, CNN-based methods can automatically extract features and do not need manual feature extraction and selection. In [21], the histogram of oriented gradients (HOG) feature was presented; it showed good results for pedestrian detection with a linear support vector machine. On the basis of HOG features, Felzenszwalb *et al.* [22] developed the mixtures of multi-scale deformable part model, which can handle pose variations of pedestrians very well. In [23], the HOG feature was further extended by adding the LUV colour feature to generate the integral channel feature (ICF). The ICF combined with boosted decision forests was very useful for pedestrian detection, outperforming previous detectors by a large margin. Then based on the ICF feature, aggregated channel features (ACFs) are proposed [24], in which new gradient magnitude channel feature is added with ACF, computational costs can be reduced by downsampling the image channels by a factor of four. To achieve better performance, a lot of studies based on HOG + LUV were proposed. SquaresChntrs [25] combined square regions from every channel for classifier training. In [26], a statistical model was employed to distinguish pedestrian body shapes. Locally decorrelated channel features (LDCF) [27] demonstrated an efficient feature transform that removed correlations in local neighbourhoods. Checkerboards [28] combined several filters including uniform squares, horizontal and vertical gradient detectors, and 'checkerboard like' patterns.

Recently, a large number of studies based on CNNs have pushed pedestrian detection results to a higher level. Contrary to the handcrafted channel-based methods, the CNN-based methods have good self-learning ability to automatically extract discriminative features. In [7], an integrated method called convolutional channel features was proposed. It trained low-level CNN features by boosting forest classifiers. To overcome the heavily occluded problem, a robust architecture deep parts [10] was proposed, which employed multiple parts for training detectors. Tian *et al.* [8] introduced a novel task-assistant CNN that used semantic information of persons and scenes for learning discriminative representations for pedestrian detection. Cai *et al.* [9] designed a complexity-aware cascade training algorithm, combining various features from both the handcrafted and the CNNs. Zhang *et al.* [6] developed RPN to generate candidates. The top performing approaches on the Caltech pedestrian benchmark are variations of fast or faster R-CNN. In [5], a multi-scale network was proposed that can detect objects in various scales well. Du *et al.* [4] proposed an fused deep neural network (F-DNN) + semantic segmentation (SS) framework that adds segmentation as a strong information for pedestrian detection.

Despite CNN-based methods making notable progress on pedestrian detection when illumination is good, the challenge of how to integrate information from multispectral data sources has been rarely investigated. In [18], a multispectral ACF detector was introduced. On the basis of ACF, it added thermal intensity channel feature T and HOG feature of the thermal image THOG, then adopted ACF + T + THOG as a desirable integration for the channel feature and adopted AdaBoost classifier for pedestrian detection. Choi *et al.* [16] introduced a CNN-based joint framework that trained end-to-end CNNs for pedestrian proposal generation in visible and thermal images individually. The final classification was performed by support vector regression on accumulated proposals. An R-CNN-based framework was presented in [17], which processed the visible and thermal data separately in R-CNN subnetworks and concatenated the two resulting features. In [14], where to fuse visible and thermal images between two streams of faster R-CNN architecture was discussed. All fusion models were based on faster R-CNN architecture, and the halfway fusion strategy which inserted fusion layer after the fourth convolutional layer showed the best performance. Fusion RPN + boosted decision tree (BDT) [15] proposed two-stream RPN architecture for proposal generation and used BDT for classification. Similar to [14], only one concatenation layer was used, leaving a lot of room for improvement. Finally, Xu *et al.* [29] proposed a cross-modality learning CNN that use multispectral images for training and the only visible image is used for testing, for which detection accuracy is not ideal. Although aforementioned efforts have been made for developing multispectral fusion algorithm for pedestrian detection, the challenge of how to fuse visible and thermal data inside CNNs for optimal performance has not been solved.

In this research, a new multi-layer fusion RPN is developed to make the most of multispectral images. It can produce accurate pedestrian candidates at various scales because multi-layer contains multiple scale features. As a result, our new MLF-CNN achieves desirable detection accuracy when compared with well-known previous methods.

3 Proposed MLF-CNN

Our architecture consists of two stages: an MLF multi-scale RPN to generate candidate proposals and a detection network that classifies these proposals using convolutional features from multiple feature maps. An overview of our MLF-CNN framework is shown in Fig. 2. The network first processes visible and thermal images with several convolutional and pooling layers to produce convolutional feature maps as well as pedestrian candidate proposals. Then, for each pedestrian proposal, ROI pooling layers extract a feature vector in fixed length from the deconvolutional feature maps. Each feature vector is sent into a fully connected layer that finally divides into two output layers: one that generates softmax probability to estimate two classes (pedestrian versus non-

pedestrian) and another layer that outputs four parameterised numbers to locate each pedestrian.

3.1 Multi-layer fused RPN

Depending on the distance between a target and the camera, the pedestrians have a wide variety of sizes. Most recent multispectral pedestrian detection efforts [14, 15] are based on RPN in faster R-CNN [30] to fuse the visible and thermal information and to generate candidate bounding boxes (The input of RPN is an image of any size, the output are a set of rectangular object proposals, each with a confidence score.). The RPN generates candidate bounding boxes from a single convolutional layer (Conv5), which is not good at handling pedestrians of various scales and is likely to miss small-sized pedestrians. To overcome the pedestrian scale diversity problem, we use a multi-scale RPN to fuse the visible and thermal information and to generate pedestrian candidates. The multi-scale RPN generates candidate bounding boxes from four layers at different scales. As shown in Fig. 2, we use the same multi-scale RPN to process both the visible image and the thermal image. Then, we have two multi-scale RPN networks, namely RPN-V and RPN-T. For each RPN, the Conv4 layer in the lower level is better for detecting small pedestrians, since it has smaller receptive fields for matching objects. In contrast, the higher-level Conv5 layer is more suitable for detecting pedestrians of large scale. According to this characteristic, we generate detection output layers from the Conv4 layer, Conv5 layer, Conv6 layer, and Pooling6 layer for both of RPN-V and RPN-T, which are denoted as Det1-V–Det4-V and Det1-T–Det4-T, respectively. For each RPN, these four output detection layers are at a different scale. We assume that the input image size is $W \times H \times D$, where W, H, and D are the width, height, and the number of channels. The detection output layers' size of Det1–Det4 are $(W/8) \times (H/8) \times c$, $(W/16) \times (H/16) \times c$, $(W/32) \times (H/32) \times c$, and $(W/64) \times (H/64) \times c$, respectively. Here, the parameter c is a sum of the number of classes and the number of bounding-box coordinates.

3.1.1 Where to fuse two layers: The current problem is to fuse RPN-V and RPN-T. The fusion can be applied at any point in the two networks, and implementing fusion at different positions of RPN would lead to different detection results. As we mentioned above, features at different layers correspond to various levels of semantic meanings and fine visual details. To fuse fine visual details at a low level as well as fuse semantic features at a high level, we insert four fusion layers at different detection output layers. As shown in Fig. 2, for both of RPN-V and RPN-T, there are a total of eight output layers Det1-T–Det4-T and Det1-V–Det4-V. The four fusion layers are inserted after the detection output layers. Then, the candidate proposals are generated from the four-fused detection output layers. This MLF-RPN can build a stronger and more accurate strategy to generate pedestrian candidates. The multi-layer fusion provides a set of variable receptive field scales that can cover an extensive range of pedestrian scales. Simultaneously, the Conv4-V layer and Conv4-T layer are integrated into a fused Conv4 layer. The generated multiple convolutional feature maps are served for the detection stage.

3.1.2 How to fuse two layers: In Section 3.1.1, we introduced an MLF-RPN to fuse two networks and to generate the pedestrian candidates. In this section, we introduce a summation algorithm to fuse two layers which can make the most of the multispectral data. The goal is to integrate the two layers I^{V} and I^{T} from RPN-V and RPN-T, respectively, to a fused layer I^{F}. Here, I^{V}, $I^{\mathrm{T}} \in \mathbb{R}^{H \times W \times D}$, where W, H, and D are the width, height, and the number of channels of the feature maps.

Previous CNN-based fusion methods [14–17] are directly along the dimensions d of the channels to concatenate two layers at the same position (i, j), which can be defined as

$$I^{\mathrm{F}} = f^{\mathrm{concat}}\left(I^{\mathrm{V}}, I^{\mathrm{T}}\right), \tag{1}$$

$$I^{\mathrm{F}}_{i,\,j,\,2d} = I^{\mathrm{V}}_{i,\,j,\,d}, \tag{2}$$

$$I^{\mathrm{F}}_{i,\,j,\,2d-1} = I^{\mathrm{T}}_{i,\,j,\,d}, \tag{3}$$

where $1 \le i \le H$, $1 \le j \le W$, $1 \le d \le D$ and $I^{\mathrm{F}} \in \mathbb{R}^{H \times W \times 2D}$.

This concatenation fusion leads to doubling the number of feature maps. To reuse the filters' weight in the pertained CNN model, the concatenation layer is then used as input to an inner product layer for dimension reduction. This additional multiplication by a matrix incurs a greater computational cost.

In this paper, we proposed another simpler way of fusing two layers. We use the summation of the two feature maps at the same position (i, j) and feature channels d

$$I^{\mathrm{F}} = f^{\mathrm{sum}}\left(I^{\mathrm{V}}, I^{\mathrm{T}}\right), \tag{4}$$

$$I^{\mathrm{F}}_{i,\,j,\,d} = I^{\mathrm{V}}_{i,\,j,\,d} + I^{\mathrm{T}}_{i,\,j,d}, \tag{5}$$

where $1 \le i \le H$, $1 \le j \le W$, $1 \le d \le D$ and $I^{\mathrm{F}} \in \mathbb{R}^{H \times W \times D}$

This summation fusion is a simple addition of the corresponding numbers in two layers. In the learning stage, the best correspondence can be automatically determined, which can optimise the filters' weight of each network to give us strong features from fused layers. Another advantage is that summation fusion does not increase the feature dimension, while the concatenation fusion doubles the dimension. When inserting the same number of fusion layers, summation fusion takes less training time and less testing time.

In the experimental section (Section 4.2.1), we evaluate and compare the performance of these two fusion methods in terms of their detection miss rate and speed.

3.2 Detection network

The MLF-RPN introduced above generates pedestrian candidate proposals as well as convolutional feature maps. In this section, we introduce the detection stage for classifying the candidate proposals based on the convolutional feature maps.

To compute features for a region proposal, we must first convert the image data in that region into a form that is compatible with the CNN since the input size of the pertained CNN has a natural scale (e.g. 224×224). Previous methods solve this problem through warping by upsampling the input patches' size. However, there are bad effects of input upsampling: large memory requirements reduced training speed and reduced testing speed. It should be noted that input upsampling does not enrich the visual details and still misses small-sized pedestrians. Instead, it is necessary because the higher convolutional layers respond very weakly to tiny pedestrians. Mapping a 48×48 pedestrian into a 6×6 patch of the Conv4 layer and a 3×3 patch of the Conv5 layer, for example. This results in limited information for 7×7 ROI pooling. To solve this problem, following [5], we use a deconvolution layer to increase the resolution of feature maps. As shown in Fig. 2, we deconvolve the three generated feature maps to get deconvolutional feature maps of higher density and higher resolution. Feature upsampling is better than the input image patches' upsampling, without incurring extra costs for computation and memory. The additional deconvolution layer significantly improves detection performance, particularly for tiny pedestrians.

Owing to the diversity of the real environment, the best features are not necessarily in the fused Conv4-2x, but also in the Conv4-2x-V or Conv4-2x-T. Since existing methods only extracted features from a single-fused feature map, they are not robust. We take features not only from one separate layer, but from all three feature maps. After getting three deconvolutional feature maps, we extract features using ROI pooling. The ROI pooling contains two steps. First, since all proposals are in different scales, the proposals from MLF-RPN are mapped into all three deconvolutional feature maps. In the second step, the ROI pooling layer uses maximum pooling to fix the dimension of features (e.g. $7 \times 7 \times 512$). After obtaining fixed features from all three feature maps, we integrated them into one fused ROI layer through the summation fusion.

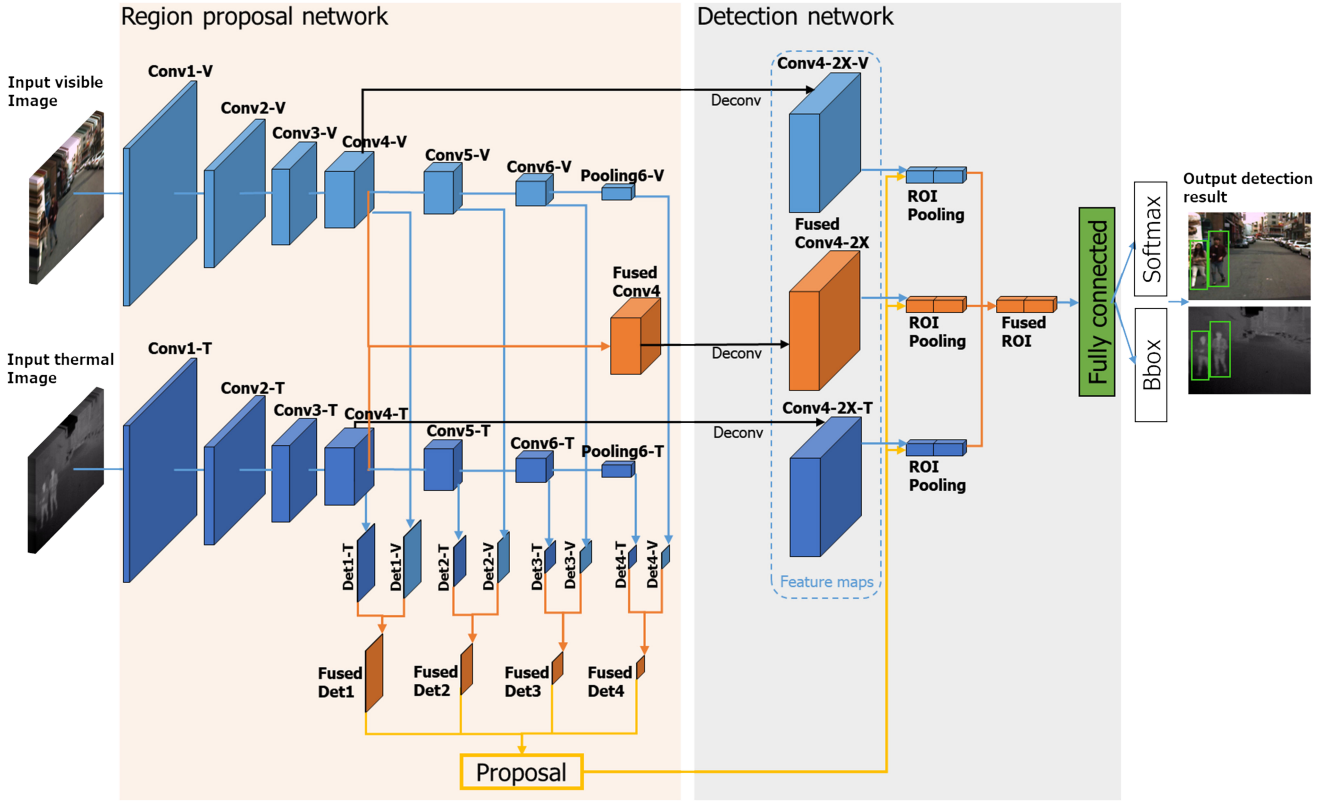


Fig. 2 *Architecture of our proposed MLF-CNN*

Finally, the features are fed to a fully connected layer, as shown in Fig. 2. Note that the ROI pooling is applied after the Conv4 layer rather than after the Conv5 layer as in [30]. This is because Conv4 corresponds to higher resolution and is more suitable for location-aware bounding-box regression. In the experimental section (Section 4.2.2), we evaluate and compare the performance of extracting features from these different layers at different resolutions.

The MLF-RPN combined with this detection network is our final MLF-CNN as shown in Fig. 2. In Section 4, we explicate our systematic experiments and analysis of the results.

3.3 Implementation details

The filters of our proposed MLF-CNN are initialised with the popular VGG-16 [31] model. Our MLF-CNN is trained end-to-end by backpropagation and stochastic gradient descent since end-to-end training can save a significant time cost [32, 30]. End-to-end training is realised by defining two loss branches: one is for the softmax classifier and the other is for linear bounding-box repressors. Therefore, the overall loss is the sum of two loss branches and optimised for multi-task. Our loss function is defined as

$$L(\{p_i\},\ \{b_i\}) = L_{\text{cls}}(p_i,\ p_i^*) + \lambda p_i^* L_{\text{reg}}(b_i,\ b_i^*), \tag{6}$$

where p_i is the probability distribution of an image patch i being a pedestrian. The value of p_i^* is ground truth label p_i^* equals to 1 if the image patch is positive, and equals to 0 if the image patch is negative. The predicted bounding box $b_i = \left(b_i^x,\ b_i^y,\ b_i^w,\ b_i^h\right)$ defining the four real-valued coordinates of the predicted bounding box and b_i^* is that of the ground truth box of a positive image patch. The value of λ is the trade-off coefficient. The classification loss L_{cls} is the log loss over pedestrian class and non-pedestrian class. The regression loss is defined as

$$L_{\text{reg}}(b_i,\ b_i^*) = \frac{1}{4} \sum_{j \in \{x,\ y,\ w,\ h\}} \text{smooth}_{L_1}(b_j,\ b_j^*), \tag{7}$$

where smooth_{L_1} is defined in [32]. The term $p_i^* L_{\text{reg}}$ means that the regression loss is activated only for $p_i^* = 1$ and is disabled otherwise.

To prevent this multi-task training instability in the early iterations, a two-stage procedure is adopted [5]. In the first stage, 10,000 iterations are run with a learning rate of 0.00005. The generated model from the first stage is used for initialising the second stage's learning. The second stage's learning rate is 0.0002 and reduced by a factor of 10 after every 10,000 iterations. The maximum number of iterations is 30,000. The parameters of layers Conv1-T, Conv1-V, Conv2-V, and Conv2-T are fixed during learning, to save training time. In addition, in order to eliminate highly overlapped bounding boxes with lower scores, non-maximum suppression is adopted after the proposal network. The value of the intersection over union (IoU) is defined as

$$\text{IoU}(b_1,\ b_2) = \frac{\text{area}(b_1 \cap b_2)}{\text{area}(b_1 \cup b_2)}, \tag{8}$$

where b_1 and b_2 are the two detection bounding boxes. In this paper, we set the threshold to 0.7 because it is experimentally demonstrated that this threshold can improve the detection efficiency without affecting the performance. The generated bounding boxes are ranked by their scores. If IoU > 0.7, it means that b_1 and b_2 highly overlap. Then the detection bounding boxes with the lower score will be eliminated.

4 Experiments

4.1 Datasets and processing platform

We evaluate our proposed fusion architecture on the KAIST multispectral pedestrian dataset [18], the UTokyo multispectral object detection dataset [19], and the OSU colour–thermal dataset [20].

The KAIST dataset [18] consists of 11 videos, which contain the aligned multispectral [red, green, and blue (RGB) visible and thermal] images, and all images are normalised to the same size of 640 × 512. In particular, the dataset contains traffic sequences with low visibility such as shadows, overexposure, dusk time, and

nighttime. The first six videos are used for training, and the remaining videos are used for testing. For training, we sample images from training sets with 2-frame skips, which contains 50,172 images (25,086 visible images and 25,086 thermal images). We evaluate on the standard 4504 images (2252 visible images and 2252 thermal images) with 20-frame skips in the test set, among which 1455 images were captured during daytime and 797 others for nighttime.

The UTokyo dataset [19] contains a total of 7512 group images (3740 taken at daytime and 3772 taken at nighttime), which were taken in a university environment at 1 fps using RGB, far infrared (FIR), mid infrared (MIR), and near infrared (NIR) cameras. Five classes (*bike*, *car*, *car_stop*, *colour_cone*, *person*) are labelled in this dataset consisting of 6066 groups of unaligned training images and 1466 groups of correctly aligned test images with a size of 320 × 256. In our case, training and testing need to use aligned images. Therefore, we only use the 1466 pair of RGB–FIR test sets for evaluation and only consider the person detection task.

The OSU colour–thermal dataset [20] has a total of 17,088 images (8544 visible images and 8544 thermal images) of the same size of 320 × 240. This dataset has a total of six sequences with each three containing scenes of the same location. Following the sampling procedure of Dollár *et al.* [33], we have uniformly sampled the frames (every 10th frame) from each of the six video sequences to include in our experiment. In total, we have 856 pairs of colour–thermal test images coming from all the six sequences.

Our processing platform is a standard personal computer with Ubuntu 14.04, with a single central processing unit (CPU) core (3.40 GHz) of an Intel Core i7-4770 with 32 GB of random access memory. An NVIDIA Titan X graphics PU was used for CNN computations. The computation environment is MATLAB R2015b.

4.2 Self-comparison of MLF-CNN

4.2.1 Comparison of fusion methods in our MLF-CNN: In this section, experimental results on different fusion methods applied to MLF-CNN are reported. We compare different fusion methods in Table 1, where we report the detection miss rate and computation time on the KAIST reasonable test set. We first observe that when using the same concatenation fusion method, the performance of our MLF-CNN is significantly better than the other two state-of-the-art methods faster R-CNN halfway fusion [14] and RPN fusion + boosted forests (BF) [15]. This indicates that our proposed MLF-CNN architecture is more effective at integrating visible and thermal information since fusion at multiple layers can not only integrate fine visual features in lower layers, but also semantic features at higher layers. Second, we see that in the MLF-CNN framework, concatenation performs considerably worst than summation fusion. Furthermore, using concatenation fusion has a much longer run time than using summation fusion. Therefore, summation fusion is an effective fusion method in our MLF-CNN framework.

Table 1 Comprehensive comparison of different fusion methods with different fusion architectures on the KAIST dataset

Fusion method	Fusion architecture	Miss rate, %	Computation time, s/f
summation	MLF-CNN (ours)	25.65	0.15
concatenation	MLF-CNN (ours)	26.77	0.20
concatenation	faster R-CNN halfway fusion [14]	37	0.19
concatenation	RPN fusion + BDT [15]	29.83	—

Table 2 Comparison of feature extraction after different layers in MLF-CNN on the KAIST dataset

Feature layer	Miss rate, %
Conv3	27.01
Conv4	25.65
Conv5	29.32

4.2.2 Comparison of feature extraction in our MLF-CNN: In this section, we conduct experiments to validate the detection performance when extracting features after different convolutional layers. Our MLF-CNN is flexible and is able to take advantage of features of high resolutions. Table 2 shows the results of extracting ROI features after different layers in our method on the KAIST reasonable test set. All entries are based on VGG-16 and the same set of MLF-RPN proposals.

Extracting ROI features after Conv4 achieves the best result of 25.65% miss rate. Extracting features after Conv5 degrades the result: the miss rate is considerably increased to 29.65%. This is because of the low-resolution features. Conv3 also shows degradation (27.01%), which can be explained by the weaker representation of the shallower layers. From this observation, we can conclude that using the Conv4 layer achieves the best performance in our MLF-CNN.

4.3 Detection evaluation on the KAIST dataset

In this section, the performance of the proposed method is compared with three other methods: (i) the KAIST baseline approach ACF + T + THOG [18], (ii) the faster R-CNN halfway fusion approach proposed in [14], and (iii) the RPN fusion + BDT approach proposed in [15]. We evaluate the detection performance using the log-average miss rate against a false-positive per image (FPPI) range of $[10^{-2}, 10^{0}]$ as suggested by Dollár *et al.* [33], and a minimum IoU threshold of 0.5 is required for a detected box to match with the ground truth box.

We first compared the detection results on a reasonable test subset, in terms of reasonable all day, reasonable daytime, and reasonable nighttime [18]. As shown in Fig. 3*a*, for reasonable all day, the RPN fusion + BDT [15], and faster R-CNN halfway fusion [14], have a respective miss rate of 29.83 and 37%, which is further reduced to as low as 25.65% by our approach. A similar trend has been observed for the other reasonable day and night subsets: for a reasonable day, we have evidenced significant improvement (over 4%) over the state-of-the-art, our approach achieves 25.22% miss rate where the top competing method does 29.58%. Meanwhile, for reasonable nighttime, the gap between ours and that of RPN fusion + BDT [15] is about 4% and the gap between ours and that of faster R-CNN halfway fusion [14] is over 10%. Furthermore, our method significantly outperforms the baseline of the ACF + T + THOG [18] method, by reducing the miss rate by about 30%. In conclusion, our method performs significantly better than all other well-known methods for detecting pedestrian instances under different illumination conditions. This demonstrates that our new architecture has advantages in fusing not only the fine visual features in lower layers to retain the visual details, but also the semantic features at higher layers, and thus can make the most use of multispectral data to detect pedestrians, even in bad illumination conditions.

We also examined detectors using three subsets of the dataset which were defined based on the size of the bounding boxes. The test sets were classified into near scale (more than 80 pixels), medium scale (30–80 px) and far scale (20–30 px) [33]. These subsets contain non-occluded pedestrians captured over the course of a day. Fig. 4*a* displays the quantitative results of the near scale. Our approach outperforms all comparison methods and achieves the lowest log-average miss rate of 12.38%, which clearly exceeds the performance of other existing methods. Furthermore, for medium scale and far scale, our approach still achieves the lowest miss rate, which amounts to substantially better performance than the existing results, as exhibited in Figs. 4*b* and *c*. In conclusion, our MLF-CNN generally outperforms other methods on the three scales. This demonstrates that our method to fuse two networks at multiple layers is more effective using multiple scales convolutional features and can detect pedestrians of various sizes.

In Table 3, we report a comparison between our method and recent multispectral pedestrian detection methods in terms of the miss rate and computational efficiency. The CPU times of all the methods were measured using the same machine. Although ACF + T + THOG [18] has very fast detection speed (0.10 s/f), the miss rate of ACF + T + THOG [18] is quite large. Therefore, MLF-CNN

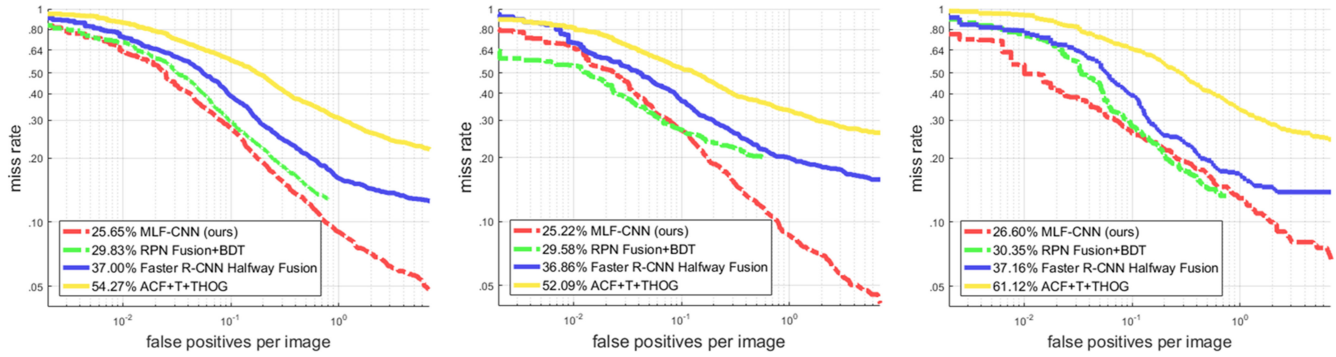


Fig. 3 *Comparison of detection results (miss rate versus FPPI) on the KAIST dataset, in terms of all day, day time, and nighttime*
(a) Reasonable all day, ***(b)*** Reasonable day, ***(c)*** Reasonable night

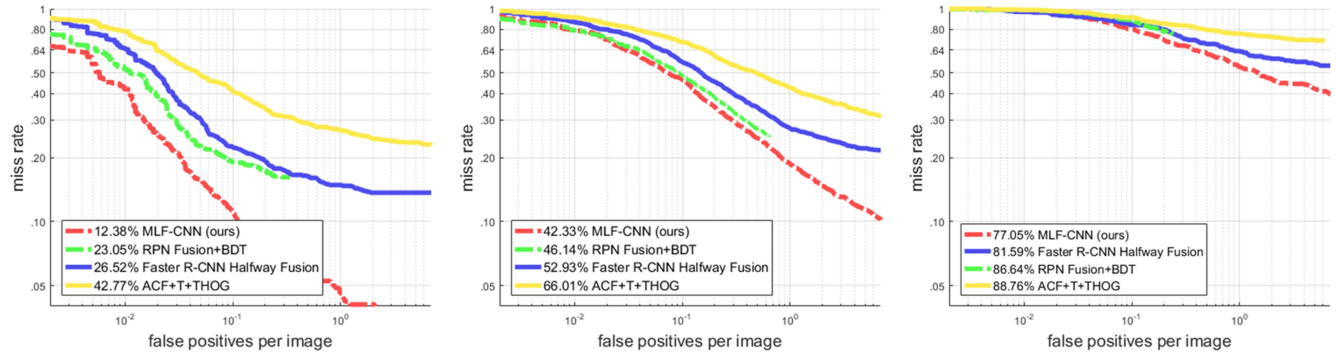


Fig. 4 *Comparison of detection results (miss rate versus FPPI) on the KAIST dataset, in terms of near scale, medium scale, and far scale*
(a) Near scale, ***(b)*** Medium scale, ***(c)*** Far scale

Table 3 Comparison of computation time and miss rate on the KAIST dataset

Method	Miss rate, %	Computation time, s/f
ACF + T + THOG [18]	54.27	0.10
faster R-CNN halfway fusion [14]	37	0.19
MLF-CNN (ours)	25.65	0.15

has a better trade-off between the detection speed and the detection performance. For testing, our network takes only 0.15 s to process one image, which is very competitive with previous methods.

To provide a visual comparison, the detection results of related works [14, 18] are illustrated and compared with our results on a set of day and night images captured in challenge scenes. As shown in Fig. 5, the first two rows show pedestrian detection in the daytime. The last three rows are nighttime pedestrian detection. In all sample images, the pedestrians are under poor illumination conditions or at various scales. It is clear that ACF + T + THOG [18] shows unsatisfactory results that produce many false-positives, while faster R-CNN halfway fusion [14] is not effective in detecting a small-scaled pedestrian. However, our proposed method works well on pedestrians at various scales both during the day and night.

4.4 Detection evaluation on the UTokyo dataset

In this section, we compare MLF-CNN with two state-of-the-art methods: ACF + T + THOG [18] and faster R-CNN halfway fusion [14]. We evaluate the performance using the log-average miss rate which is computed by averaging the miss rate at false-positive rates spaced evenly between the $[10^{-2}, 10^{0}]$ FPPI range. The comparison results are evaluated for pedestrian instances of the overall case [33], which contains all scales and occlusions.

As displayed in Fig. 6, our MLF-CNN significantly outperforms the other two state-of-the-art methods and achieves the lowest miss rate of 27.63%. Fig. 7 presents example detection results of our approach on UTokyo test images. It also provides visual comparisons, where evidently the state-of-the-art methods faster R-CNN halfway fusion [14] and ACF + T + THOG [18] produce more false alarms as well as more misses. The UTokyo dataset is challenging, but suitable for validating detection performance since the number of nighttime cases dominates the overall pedestrian population. This comparison on the UTokyo dataset indicates that our MLF-CNN is more effective to fuse visible and thermal information so that pedestrians can still be detected even in poor illumination conditions. Moreover, our method shows better generalisation ability than other methods.

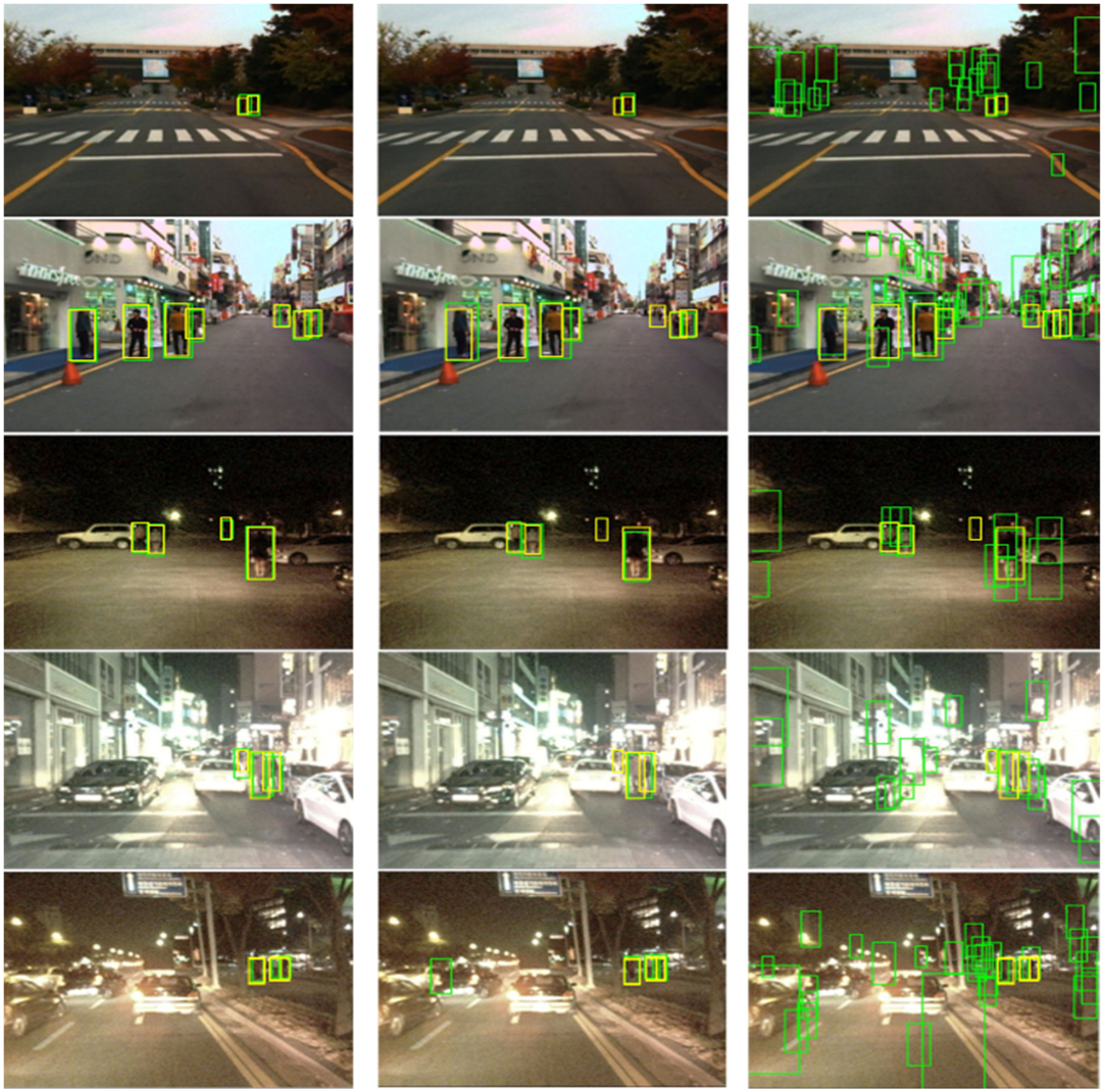
In Table 4, we make a comprehensive comparison between our method and state-of-the-art methods. The CPU times of all the methods were measured using the same machine. Although ACF + T + THOG [18] has a very fast detection speed (0.03 s/f), the miss rate of ACF + T + THOG [18] is quite large. Therefore, MLF-CNN has a better trade-off between detection speed and detection performance.

4.5 Detection evaluation on the OSU colour–thermal dataset

In this section, we evaluate the performance of the OSU colour–thermal dataset. The comparison methods and results evaluation settings are the same as the statement in Section 4.4.

Fig. 8 displays the quantitative results of the overall case. A similar trend to what we have observed for the UTokyo dataset also occurs here: the overall performance gap is quite large, 12.78% of ours versus 32.79% of the faster R-CNN halfway fusion [14]. Fig. 8 shows that our method significantly outperforms other methods.

Fig. 9 displays several visual examples with side-by-side comparisons. Faster R-CNN halfway fusion [14] and ACF + T + THOG [18] again produce considerably more false alarms and missing instances than those of our approach. This result makes sense because the OSU colour–thermal dataset involves the presence of many small-sized pedestrians and a low resolution, which demonstrate that our multi-layer fusion strategies can not



a b c

Fig. 5 *Examples of detection results at day and night on the KAIST dataset. The yellow bounding boxes denote the ground truth. The green bounding boxes show the detection results, illustrated in visible images*

(a) Detection results of MLF-CNN (ours), ***(b)*** Detection results of faster R-CNN halfway fusion [14], ***(c)*** Detection results of ACF + T + THOG [18]

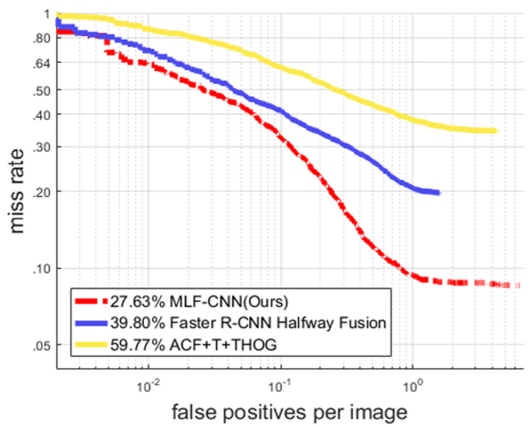


Fig. 6 *Comparison of detection results on the UTokyo dataset*

Fig. 7 *Examples of detection results on the UTokyo dataset. The yellow bounding boxes denote the ground truth, displayed in thermal images. The green bounding boxes show the detection results, illustrated in visible images*

(a) Input thermal images with ground truth, ***(b)*** Detection results of MLF-CNN (ours), ***(c)*** Detection results of faster R-CNN halfway fusion [14], ***(d)*** Detection results of ACF + T + THOG [18]

Table 4 Comparison of computation time and miss rate on the UTokyo dataset

Method	Miss rate, %	Computation time, s/f
ACF + T + THOG [18]	59.77	0.03
faster R-CNN halfway fusion [14]	39.80	0.13
MLF-CNN (ours)	27.63	0.08

only cover pedestrians of various scales, but also has better robustness.

Table 5 compares the detection speed of the proposed MLF-CNN with those of other reported methods. We can see that MLF-CNN achieves a better balance between speed and accuracy.

5 Conclusions and future works

In this paper, we proposed a novel and efficient MLF-CNN for multispectral pedestrian detection in adverse illumination scenes. Our MLF-CNN contains two stages: a region proposal stage and a detection stage. In the region proposal stage, we designed a multi-layer fusion RPN to effectively fuse the visible and thermal information and proposed to use summation fusion to integrate two convolutional layers. This multi-layer fusion RPN makes the most use of multispectral images and generates accurate pedestrian candidates in various scales. In the detection stage, we extracted features from three feature maps and combined all features through the fused ROI pooling layer. Using the fused ROI pooling layer for classification further improved the detector robustness under diverse environments.

Experiments at different settings on our multispectral pedestrian dataset have shown that our proposed MLF-CNN works well under various illumination conditions to detect pedestrians at different scales. It outperforms all previous methods, reducing detection miss rate by 4.18% on reasonable all-day test sets when compared with the previous state-of-the-art method. Moreover, our MLF-CNN detector achieves competitive computation speed of about 7 fps. We believe that our work will make a valuable contribution to the area of pedestrian detection using a fusion approach.

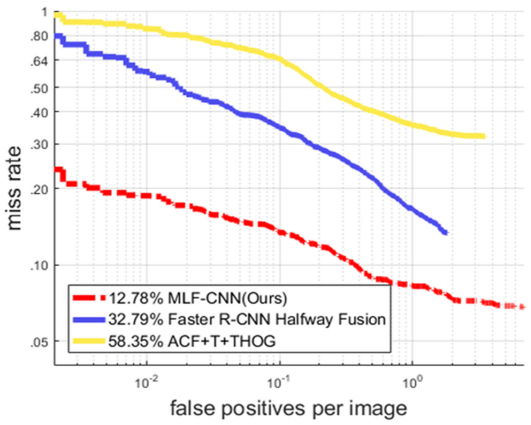


Fig. 8 *Comparison of the detection results on the OSU colour–thermal dataset*

In the future, we plan to develop a cascaded classifier trained by using CNN features in the detection stage. The cascade structure can reject many negative background samples early saving computational cost. Furthermore, sample re-weighting in the cascade structure can be helpful in reducing false negative cases. Furthermore, we plan to develop an adaptive weighting mechanism to better fuse the visible and thermal information under diverse illumination conditions of the input multispectral images.

6 Acknowledgments

This material is based on work supported by the Ministry of Trade, Industry & Energy (MOTIE, Korea) under Industrial Technology Innovation Program (10080619).

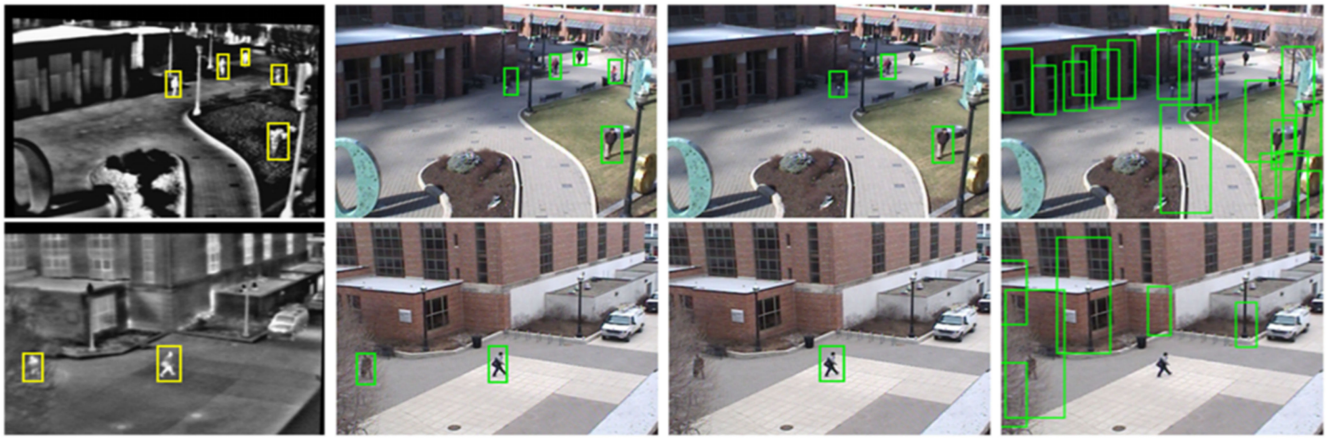


Fig. 9 *Examples of detection results on the OSU colour–thermal dataset. The yellow bounding boxes denote the ground truth, displayed in thermal images. The green bounding boxes show the detection results, illustrated in visible images*
(a) Input thermal images with ground truth, ***(b)*** Detection results of MLF-CNN (ours), ***(c)*** Detection results of faster R-CNN halfway fusion [14], ***(d)*** Detection results of ACF + T + THOG [18]

Table 5 Comparison of computation time and miss rate on the OSU colour–thermal dataset

Method	Miss rate, %	Computation time, s/f
ACF + T + THOG [18]	58.35	0.028
faster R-CNN halfway fusion [14]	32.79	0.152
MLF-CNN (ours)	12.78	0.069

7 References

[1] Wang, X., Wang, M., Wei, L.: 'Scene-specific pedestrian detection for static video surveillance', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2014, **36**, (2), pp. 361–374

[2] Xiaokai, L.: 'Pedestrian re-identification via coarse-to-fine ranking', *IET Comput. Vis.*, 2016, **10**, (5), pp. 366–373

[3] Tang, S., Andriluka, M., Schiele, B.: 'Detection and tracking of occluded people', *Int. J. Comput. Vis.*, 2014, **110**, (1), pp. 58–69

[4] Du, X., El-Khamy, M., Lee, J., *et al.*: 'Fused DNN: a deep neural network fusion approach to fast and robust pedestrian detection'. Proc. IEEE Winter Conf. Applications of Computer Vision, Santa Rosa, CA, USA, March 2017, pp. 953–961

[5] Cai, Z., Fan, Q., Feris, R.S., *et al.*: 'A unified multi-scale deep convolutional neural network for fast object detection'. Proc. European Conf. Computer Vision, Amsterdam, The Netherlands, October 2016, pp. 354–370

[6] Zhang, L., Lin, L., Liang, X., *et al.*: 'Is faster R-CNN doing well for pedestrian detection?'. Proc. European Conf. Computer Vision, Amsterdam, The Netherlands, October 2016, pp. 443–457

[7] Yang, B., Yan, J., Lei, Z., *et al.*: 'Convolutional channel features'. Proc. IEEE Int. Conf. Computer Vision, Santiago, Chile, December 2015, pp. 82–90

[8] Tian, Y., Luo, P., Wang, X., *et al.*: 'Pedestrian detection aided by deep learning semantic tasks'. Proc. IEEE Conf. Computer Vision and Pattern Recognition, Boston, MA, USA, June 2015, pp. 5079–5087

[9] Cai, Z., Saberian, M., Vasconcelos, N.: 'Learning complexity-aware cascades for deep pedestrian detection'. Proc. IEEE Int. Conf. Computer Vision, Santiago, Chile, December 2015, pp. 3361–3369

[10] Tian, Y., Luo, P., Wang, X., *et al.*: 'Deep learning strong parts for pedestrian detection'. Proc. IEEE Int. Conf. Computer Vision, Santiago, Chile, December 2015, pp. 1904–1912

[11] Ngiam, J., Khosla, A., Kim, M., *et al.*: 'Multimodal deep learning'. Proc. Int. Conf. Machine Learning, Bellevue, Washington, USA, June 2011, pp. 689–696

[12] Feichtenhofer, C., Pinz, A., Zisserman, A.P.: 'Convolutional two-stream network fusion for video action recognition'. Proc. IEEE Conf. Computer Vision and Pattern Recognition, Las Vegas, NV, USA, June 2016, pp. 1933–1941

[13] Wang, L., Li, Y., Lazebnik, S.: 'Learning deep structure-preserving image-text embeddings'. Proc. IEEE Conf. Computer Vision and Pattern Recognition, Las Vegas, NV, USA, June 2016, pp. 5005–5013

[14] Liu, J., Zhang, S., Wang, S., *et al.*: 'Multispectral deep neural networks for pedestrian detection'. Proc. British Machine Vision Conf., York, UK, September 2016, pp. 1–13

[15] König, D., Adam, M., Jarvers, C., *et al.*: 'Fully convolutional region proposal networks for multispectral person detection'. Proc. IEEE Workshop on Computer Vision and Pattern Recognition, Honolulu, HI, USA, July 2017, pp. 243–250

[16] Choi, H., Kim, S., Park, K., *et al.*: 'Multi-spectral pedestrian detection based on accumulated object proposal with fully convolutional networks'. Proc. IEEE Int. Conf. Pattern Recognition, Cancun, Mexico, December 2016, pp. 621–626

[17] Wagner, J., Fischer, V., Herman, M., *et al.*: 'Multispectral pedestrian detection using deep fusion convolutional neural networks'. Proc. European Symp. Artificial Neural Networks, Bruges, Belgium, April 2016, pp. 509–514

[18] Hwang, S., Park, J., Kim, N., *et al.*: 'Multispectral pedestrian detection: benchmark dataset and baseline'. Proc. IEEE Conf. Computer Vision and Pattern Recognition, Boston, MA, USA, June 2015, pp. 1037–1045

[19] Takumi, K., Watanabe, K., Ha, Q., *et al.*: 'Multispectral object detection for autonomous vehicles'. Proc. Thematic Workshops of ACM Multimedia, Mountain View, CA, USA, October 2017, pp. 35–43

[20] Davis, J.W., Sharma, V.: 'Background-subtraction using contour-based fusion of thermal and visible imagery', *Comput. Vis. Image Underst.*, 2007, **106**, (2–3), pp. 162–182

[21] Dalal, N., Triggs, B.: 'Histograms of oriented gradients for human detection'. Proc. IEEE Conf. Computer Vision and Pattern Recognition, San Diego, CA, USA, June 2005, pp. 886–893

[22] Felzenszwalb, P., McAllester, D., Ramanan, D.: 'A discriminatively trained, multiscale, deformable part model'. Proc. IEEE Conf. Computer Vision and Pattern Recognition, Anchorage, AK, USA, June 2008, pp. 1–8

[23] Dollár, P., Tu, Z., Perona, P., *et al.*: 'Integral channel features'. Proc. British Machine Vision Conf., London, UK, September 2009, pp. 1–11

[24] Dollár, P., Appel, R., Belongie, S., *et al.*: 'Fast feature pyramids for object detection', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2014, **36**, (8), PP. pp. 1532–1545

[25] Benenson, R., Omran, M., Hosang, J., *et al.*: 'Ten years of pedestrian detection, what have we learned?'. Proc. ECCV Workshop Vision for Road Scene Understanding and Autonomous Driving, Zurich, Switzerland, September 2014, pp. 613–627

[26] Zhang, S., Bauckhage, C., Cremers, A.B.: 'Informed Haar-like features improve pedestrian detection'. Proc. IEEE Conf. Computer Vision and Pattern Recognition, Columbus, OH, USA, June 2014, pp. 947–954

[27] Woonhyun, N., Dollár, P., Hee, H.J.: 'Local decorrelation for improved detection'. Proc. Int. Conf. Neural Information Processing Systems, Montreal, Canada, December 2014, pp. 424–432

[28] Zhang, S., Benenson, R., Schiele, B.: 'Filtered channel features for pedestrian detection'. Proc. IEEE Conf. Computer Vision and Pattern Recognition, Boston, MA, USA, June 2015, pp. 1751–1760

[29] Xu, D., Ouyang, W., Ricci, E., *et al.*: 'Learning cross-modal deep representations for robust pedestrian detection'. Proc. IEEE Conf. Computer Vision and Pattern Recognition, Honolulu, HI, USA, July 2017, pp. 4236–4244

[30] Ren, S., He, K., Girshick, R., *et al.*: 'Faster R-CNN: towards real-time object detection with region proposal networks', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2017, **39**, (6), pp. 1137–1149

[31] Simonyan, K., Zisserman, A.: 'Very deep convolutional networks for large-scale image recognition'. Proc. Int. Conf. Learning Representations, San Diego, CA, May 2015, pp. 1–14

[32] Girshick, R.: 'Fast R-CNN'. Proc. IEEE Int. Conf. Computer Vision, Santiago, Chile, December 2015, pp. 1440–1448

[33] Dollár, P., Wojek, C., Schiele, B., *et al.*: 'Pedestrian detection: an evaluation of the state of the art', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2012, **34**, (4), pp. 743–761