Multi-spectral Pedestrian Detection Based on Accumulated Object Proposal with Fully Convolution Network

Hangil Choi, Seungryong Kim, Kihong Park, and Kwanghoon Sohn
School of Electrical and Electronic Engineering, Yonsei University, Seoul, Korea
Email: khsohn@yonsei.ac.kr

***Abstract*—This paper presents a method for detecting a pedestrian by leveraging multi-spectral image pairs. Our approach is based on the observation that a multi-spectral image, especially far-infrared (FIR) image, enables us to overcome inherent limitations for pedestrian detection under challenging circumstances, such as even dark environments. For that task, multi-spectral color-FIR image pairs are used in a synergistic manner for pedestrian detection through deep convolutional neural networks (CNNs) learning and support vector regression (SVR). For inferring the confidence of a pedestrian, we first learn end-to-end CNNs between color images (or FIR images) and bounding box annotations of pedestrians, respectively. Furthermore, for each object proposal, we extract intermediate activation features from network, and learn the probability of pedestrian using SVR. To improve the detection performance, proposal-wise pedestrian probabilities are accumulated on the image domain. Based on the pedestrian confidence estimated from each network and accumulated pedestrian probabilities, the most probable pedestrian is finally localized among object proposal candidates. Thanks to its high robustness of multi-spectral imaging in dark environments and its high discriminative power of deep CNNs, our framework is shown to surpass state-of-the-art pedestrian detection methods on multi-spectral pedestrian benchmark.**

I. Introduction

Pedestrian detection is one of the most extensively studied research fields in many computer vision applications, such as surveillance and intelligent vehicle systems. Although a variety of methods have been proposed [1], [2], [3] for a long time, accurate and robust pedestrian detection is still regarded as a challenging task due to its inherent limitations such as intra-class variations, tiny appearances, and bad visibility at night, hindering its application to practical systems [8]. More specifically, as shown in Fig. 1, it is difficult to distinguish pedestrians wearing clothes of various colors and shapes in Fig. 1(a). It is also hard to detect pedestrians located far from the camera due to their small portion of area in an image in Fig. 1(b). Moreover, at night environment, the visibility of a pedestrian is limited due to non-uniform or poor illumination conditions in Fig. 1(c).

To alleviate these problems, a number of approaches have been proposed [4], [16], [2]. Conventionally, many methods have utilized hand-crafted features from images, such as Haar [4], Local Binary Pattern (LBP) [16], and Histogram of Oriented Gradients (HOG) [2], and classifiers, such as, Support Vector Machine (SVM) [10], or AdaBoost [3]. As a primary work using these hand-crafted feature, the deformable part based model (DPM) [5] that uses HOG features, and the aggregated channel feature (ACF) [3] that combines LUV color channels, gradient magnitude, and histogram, have been widely used, but these methods also have limitations coming from low discriminative power of their features in challenging circumstances as shown in Fig. 2. Recently, deep convolutional neural networks (CNNs) achieved substantially promising results upon the state-of-the-art in pedestrian detection, owing to their capacity to represent discriminative features from raw pixels [15], [17]. However, although many CNN-based frameworks provide robustness in day circumstances, they have a fatal difficulty at night due to poor illumination [8].

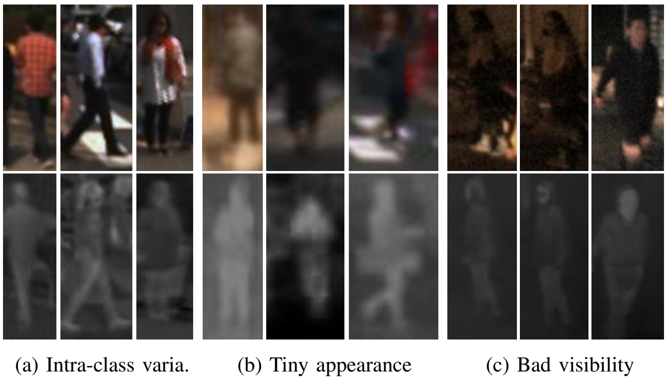

(a) Intra-class varia. (b) Tiny appearance (c) Bad visibility

Fig. 1. Examples of multi-spectral (top) RGB and (bottom) FIR images for visualizing challenging limitations in practical pedestrian detection, such as (a) intra-class variation, (b) tiny appearance, and (c) bad visibility at night. By simultaneously leveraging RGB and FIR images, our approach solves these challenging limitations in pedestrian detection.

Recently, an infrared spectrum image has provided alternative information in many computer vision and computational photography applications to provide additional information [9]. Especially, the far-infrared (FIR) camera captures the radiated heat of objects in the scene, which provides supplement information at night environments. In a similar manner, our approach starts from the intuition that the additional spectrum images, especially FIR, can be used to improve the visibility of color images in challenging conditions such as bad visibility at night circumstances. In fact, many conventional methods [9] have tried to detect a pedestrian using FIR spectrum images, but they do not employ visible and FIR spectrum domain simultaneously, rather only use FIR spectrum image.

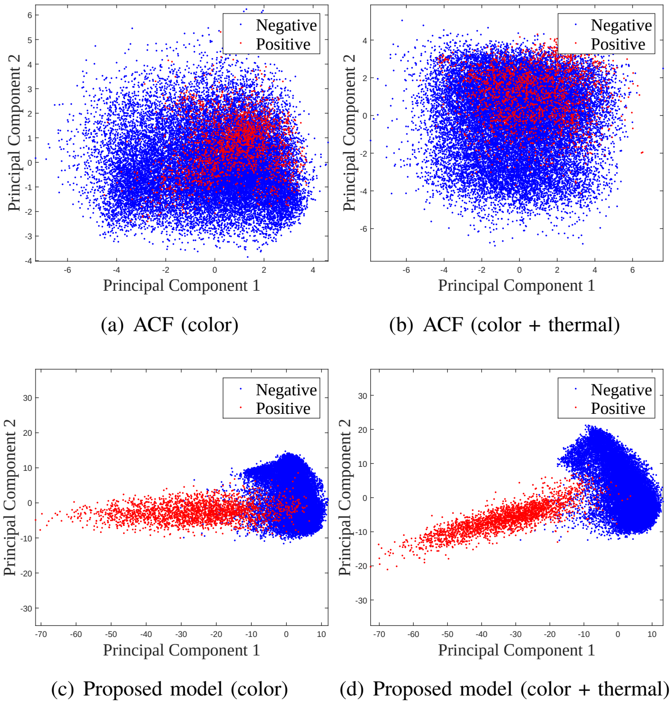
It is because it might be not an easy task to combine these multi-spectral images into one specific task due to non-linear relationship between them.

In this paper, we propose joint framework for robust pedestrian detection under challenging day and night circumstances using multi-spectral RGB and FIR spectrum images, by leveraging the deep convolutional neural networks (CNNs) [7] and support vector regression (SVR) [10]. For regressing the confidence of pedestrians, we first learn end-to-end CNN networks for pedestrian confidences in RGB and FIR individually to deal with a non-linear relationship. By fusing them independently trained from networks, the final pedestrian confidences can be estimated. Since we train our model by using only bounding box annotations, there might exist difficulties in finding a detailed pedestrian response. To alleviate this, we adopt accumulating proposal-wise constraints to enrich our fine-detailed pedestrian detection. Finally, in accumulated object proposal combined with independently estimated confidences from RGB and FIR networks, we localize the pedestrian using weighted summation. By simultaneously considering RGB and FIR images, our framework provides reliable pedestrian detection performance even under challenging images, which is not possible when using only RGB images. In the experimental results, we show that our pedestrian detection framework outperforms existing approaches on multi-spectral database.

To summarize, the contribution of this paper are threefold: (1) To the best of our knowledge, our approach is the first attempt to combine different spectral images, i.e., visible and infrared spectrum image, into a single framework for pedestrian detection. (2) We propose learning discriminative confidence for the pedestrian detection in a deep learning framework using end-to-end learning, i.e., fully convolution network (FCN) [7]. In order to exactly localize a pedestrian, we propose accumulating the probabilities of pedestrians using SVR [10], which providing highly improved pedestrian detection performance. (3) We experimentally demonstrate that proposed method outperforms the conventional methods, not only day image, but also night image.

II. Related Work

Hand-craft feature based approaches: Viola and Jones (VJ) [4] and HOG-based detector [2] are two representative works in the pedestrian detection. VJ [4] has a very fast detection speed since it uses simple Haar-like features and cascade of boosted classifiers. This framework is further developed by substituting simple Haar-like features into multiple types of features, which named by aggregated channel features (ACF) [3]. Although ACF [3] has a great performance under limited circumstances, it cannot have enough discriminative power in a challenging situations as shown in Fig. 2. The deformable part based model (DPM) [5] made a breakthrough in a pedestrian detection by efficiently aggregating local templates for each body parts. However, these hand-crafted features had a difficulty in finding a small pedestrian due to low discriminative power. Recently, many methods tried to use multi-spectral images to solve the inherent limitations [14]. Multispectral ACF [3] tries to solve bad visibility problem during both and night as shown in Fig. 1(c). To handle the multi-spectral images effectively, multispectral ACF [8] regards FIR as an additional one in RGB. However, it shows the limitations on challenging cases, such as far scale detection as shown in Fig. 1(b) and all day long detection. The unsatisfactory results of multispectral ACF [8] mainly come from the hand-crafted feature that does not have enough discriminative power to represent a pedestrian in challenging circumstances. In this manner, the multispectral ACF [8] cannot provide an optimal solution for all-day-long the pedestrian detection.



(a) ACF (color)

(b) ACF (color + thermal)

(c) Proposed model (color)

(d) Proposed model (color + thermal)

Fig. 2. Examples of principal component analysis (PCA) on color and both color and thermal channel features. Blue and red dots indicate negative and positive samples. Proposed model extract features at the last convolutional layer (conv5). Compared with ACF [3], which was the state-of-the-art hand-crafted method in a pedestrian detection, the features in proposed method is more discriminative. By adding additional thermal channel, the positive and negative channels become more discriminant.

Deep CNN features based approaches: In the field of pedestrian detection, CNN-based models have popularly proposed in very recently. These deep-based models were well designed to perform a pedestrian detection by modeling a specific role, such as performing layer jointly similar to DPM in JointDeep [15] and preventing confusions between human body and background in TA-CNN [17]. In the success of deep-based models, many researchers have applied various CNN-based architecture in pedestrian detection, such as Alexnet-net and VGG-net [11]. Moreover, Fully Convolutional Network (FCN) [7], which is an end-to-end learning method, has been used in semantic segmentation, predicting object labels of the whole images. However, when even CNN-based framework is applied to the pedestrian detection, they also have limitations derived from the inherent problems as shown in Fig. 1 (c).

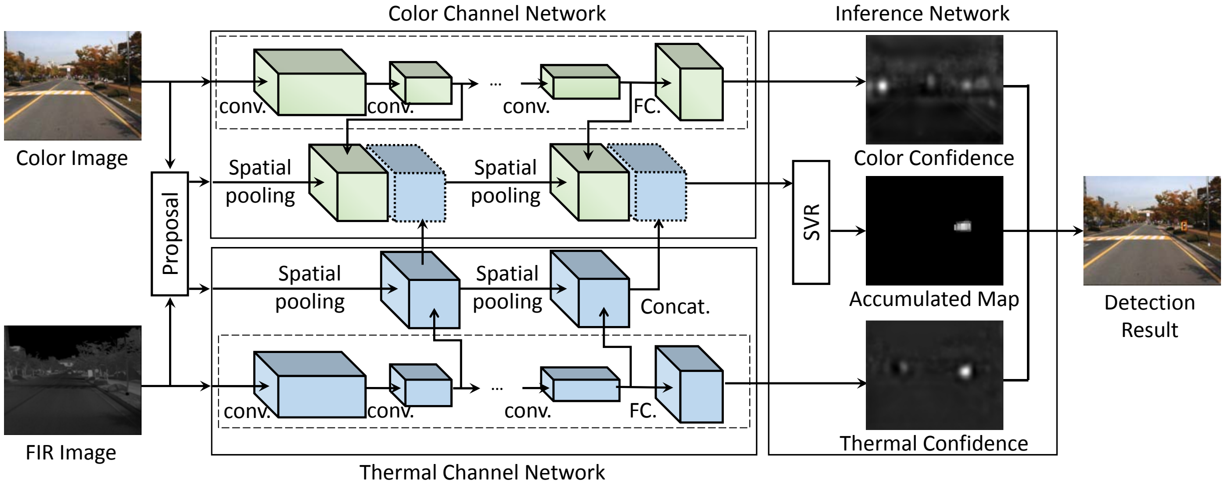


Fig. 3. The illustration of overview pipeline for pedestrian detection. Proposed framework takes RGB-FIR image pair as the input and pedestrian locations as the output. In the result images, estimated pedestrian (yellow) and ground truth (red) are described, respectively.

III. MOTIVATION

In order to observe the benefits using additional FIR channel, we compared the distributions of features of RGB channel, and RGB+FIR channels. We used multi-spectral images in KAIST benchmark [8] as shown in Fig. 2. We extracted the state-of-the-art feature method, ACF, without any technique, such as hard negative mining or jitter the data. We extracted our feature in the last convolution layer (i.e., conv5) of our CNNs, which will be explained in details.

In comparison with the ACF and CNN features, we could derive two key advantages in our framework. First, ACF feature shows the limited performance of the pedestrian detection due to its inherently limited robustness of the hand-craft features, whereas proposed feature has more discriminative power derived from deep CNN as shown in Fig. 2. Second, the joint use of RGB and FIR channel efficiently make pedestrian features more discriminative regardless of day and night.

IV. PROPOSED METHOD

A. Problem statement and overview

Given a color image $I_i : \mathcal{I} \rightarrow \mathbb{R}^3$ and a thermal image $F_i : \mathcal{I} \rightarrow \mathbb{R}$ pairs for pixel $i = [x_i, y_i]$, where $\mathcal{I} \subset \mathbb{N}^2$ is a discrete image domain, our aim is to robustly localize the location of pedestrian under challengingly varying conditions such as day and night circumstances. To realize this task, our approach is to choose the most probable object proposal from a number of object proposal candidates $t^k = (t^k_x, t^k_y, t^k_w, t^k_h)$ with locations t^k_x, t^k_y and size t^k_w, t^k_h for $k = 1, ..., N_{\text{oc}}$ with the number of object proposal candidates N_{oc}. Note that they can be estimated using any existing object proposal methods. With object proposals and our trained end-to-end CNN network, we extract pedestrian features by considering RGB and FIR features simultaneously. To enrich our pedestrian confidence fine-detailed, we combine the individual proposal-wise predictions via greedy strategy applied to accumulate proposal-wise pedestrian probability with jointly trained SVR.

B. Pedestrian confidence estimation using FCN

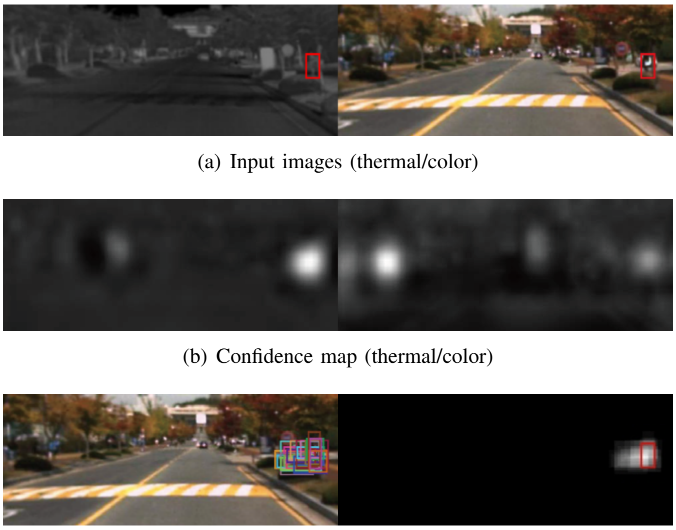
First of all, for regressing the probability of a pedestrian, we use end-to-end FCN learning [7] by independently learning as shown in Fig. 3. These two models defined by $\mathcal{F}(I_i; \mathbf{w}^I_\mathcal{F})$ and $\mathcal{F}(F_i; \mathbf{w}^F_\mathcal{F})$, where $\mathbf{w}^I_\mathcal{F}$ and $\mathbf{w}^F_\mathcal{F}$ represent network parameters for RGB and FIR images. These networks are learned between RGB image I_i and ground truth bounding box annotation for pedestrian P_i, and FIR image F_i and P_i, respectively. Using the feed-forward process from these networks, we estimate the probability of pedestrian $\mathcal{C}^I = \mathcal{F}(I_i; \mathbf{w}^I_\mathcal{F})$ from RGB image, and the probability of pedestrian $\mathcal{C}^F = \mathcal{F}(F_i; \mathbf{w}^F_\mathcal{F})$ from FIR image, respectively. However, since each domain has inherent limitations (e.g., RGB image is sensitive to the light, while FIR image provides high noise and a lack of texture), these independent learning framework cannot provide reliable detection performance. Thus, with $\mathcal{C}^I$ and $\mathcal{C}^F$, our approach uses joint intermediate feature responses from each network to reliably detect pedestrian.

C. Feature extraction from intermediate activations

Pedestrian detection task has inherent limitations as described in Fig. 1. To alleviate these limitations, our approach starts from the assumption that the average features of intermediate activations (i.e., conv5) for all proposals have the sufficient discriminative power. It is shown that extracting features from the last intermediate activations is robust to object deformations according to [6]. However, when the layer is going deeper, the discriminative power of features increases, but its resolution decrease at the same time. To address this, we extract features via multiple intermediate activations.

Specifically, to formulate pedestrian detection using RGB and FIR images in a single task, we concatenate intermediate activations (e.g., conv5) computed from RGB and FIR network $A_i \in \mathbb{R}^{w \times h \times d}$. Note that spatial resolution $w \times h$ are smaller than input image size $W \times H$. Thus, we aggregate the features within each object proposal t^k such that

$$\hat{A}^k = Z^{-1} \sum\nolimits_{i \in \Omega^k} A_i, \tag{1}$$



(a) Input images (thermal/color)

(b) Confidence map (thermal/color)

(c) Positive proposals with joint SVR

(d) Accumulated map

Fig. 4. Visualization on the impact of accumulated object proposal using SVR. In input images, red denotes ground truth. Confidence map on RGB and FIR show different score, respectively. However, with accumulated proposal-wise probability map using joint SVR, we localize a pedestrian more accurately. The results of the accumulated map in (d) are more fine-detailed than independently learned confidence map in (b).

where the constraint $\Omega^k = [\Omega_x^k, \Omega_y^k]$ is defined such that $\lfloor w/W \rfloor t_x^k \leq \Omega_x^k \leq \lfloor w/W \rfloor (t_x^k + t_w^k)$, $\lfloor h/H \rfloor t_y^k \leq \Omega_y^k \leq \lfloor h/H \rfloor (t_y^k + t_h^k)$, and normalization term $Z = \lfloor w/W \rfloor t_w^k + \lfloor h/H \rfloor t_h^k$. The annotation $\lfloor \cdot \rfloor$ denote floor operation.

By using concatenated activation feature $\hat{A}^k$, we estimate the probability of pedestrian on each object proposal.

D. Ensemble with accumulated object proposal using SVR

The concatenated activation feature $\hat{A}^k$ might be included distinct features as shown in Fig. 4 (b). Because they trained independently to preserve inherent characteristic. In order to activate $\hat{A}^k$ correctly, we re-train support vector regression modeling multiple activation for prevent the confusing response and accumulating proposal P^k to estimate dense probability map C as described in Fig. 4(c-d).

Specifically, the accumulate probability map $\mathcal{C}^A$ can computed by considering all k and i such that

$$\mathcal{C}^A = K^{-1} \sum_{k \in \{1,...,N_{\mathrm{oc}}\}} \sum_{i \in \Omega^k} P_i^k, \tag{2}$$

where P_i^k is a pixel-wise probability P^k of object proposal t^k. Normalization term K is the number of object proposals existing in Ω^k.

E. Pedestrian localization using probability ensemble

To enrich the details of a pedestrian detection, we combine the accumulated probability map $\mathcal{C}^A$ and probability $\mathcal{C}^I$ and $\mathcal{C}^F$ from each network. To localize a pedestrian, we choose the most probable object proposal from a number of object proposal candidates t^k. We first aggregate the probability $\mathcal{C}^A$, $\mathcal{C}^I$, and $\mathcal{C}^F$ on each object proposal candidates t^k. For $\mathcal{C}_k^A$, it can be derived such that $\mathcal{C}_k^A = \sum_{i \in \Omega^k} \mathcal{C}_i^A$. Here, $\mathcal{C}_k^I$, and $\mathcal{C}_k^F$ are similar defined.

Finally, a pedestrian location is estimated by choosing the most probable object proposal $\hat{k}$ from $k \in \{1, ..., N_{\mathrm{oc}}\}$ from weighted average probability such that

$$\hat{k} = \mathrm{argmin}_k (\alpha \mathcal{C}_k^A + \beta \mathcal{C}_k^I + \gamma \mathcal{C}_k^F), \tag{3}$$

where α, β, and γ are weight parameters. For relaxing the constraint for pedestrian detection problem, the probable object proposal set $\hat{k'}$ can be derived such that

$$\hat{k'} = \{k | \alpha \mathcal{C}_k^A + \beta \mathcal{C}_k^I + \gamma \mathcal{C}_k^F \geq \tau\}, \tag{4}$$

where τ is the threshold parameter.

V. Implementation Details

Network Configuration: Instead of a random initialization, we initialize the weights $\mathbf{w}_{\mathcal{F}}^I$ and $\mathbf{w}_{\mathcal{F}}^F$ using VGG 16-layer net pre-trained on ILSVRC dataset [11]. Our network architecture follows initial setup of FCN [7] such that the network contains 5 stages of convolutional layers, and 3 stages of fully connected layers.

Object Proposal: There exist many algorithms to generate object proposals. Among them, we employ edge-box [13] because of its efficiency and effectiveness. For testing image, we generate the object proposals RGB and FIR, respectively.

Batch Normalization: In order to optimize our RGB and FIR images, we put the batch normalization layer after every convolution layer to reduce the internal-covariate-shift. We also remove the drop-outs in fully connected layers as suggested in [12].

Optimization: We implement the proposed networks using MatConvNet library [11]. The standard stochastic gradient descent with momentum is employed for optimization, where initial learning rate, momentum and weight decay are set to 0.001, 0.9, and 0.00005, respectively.

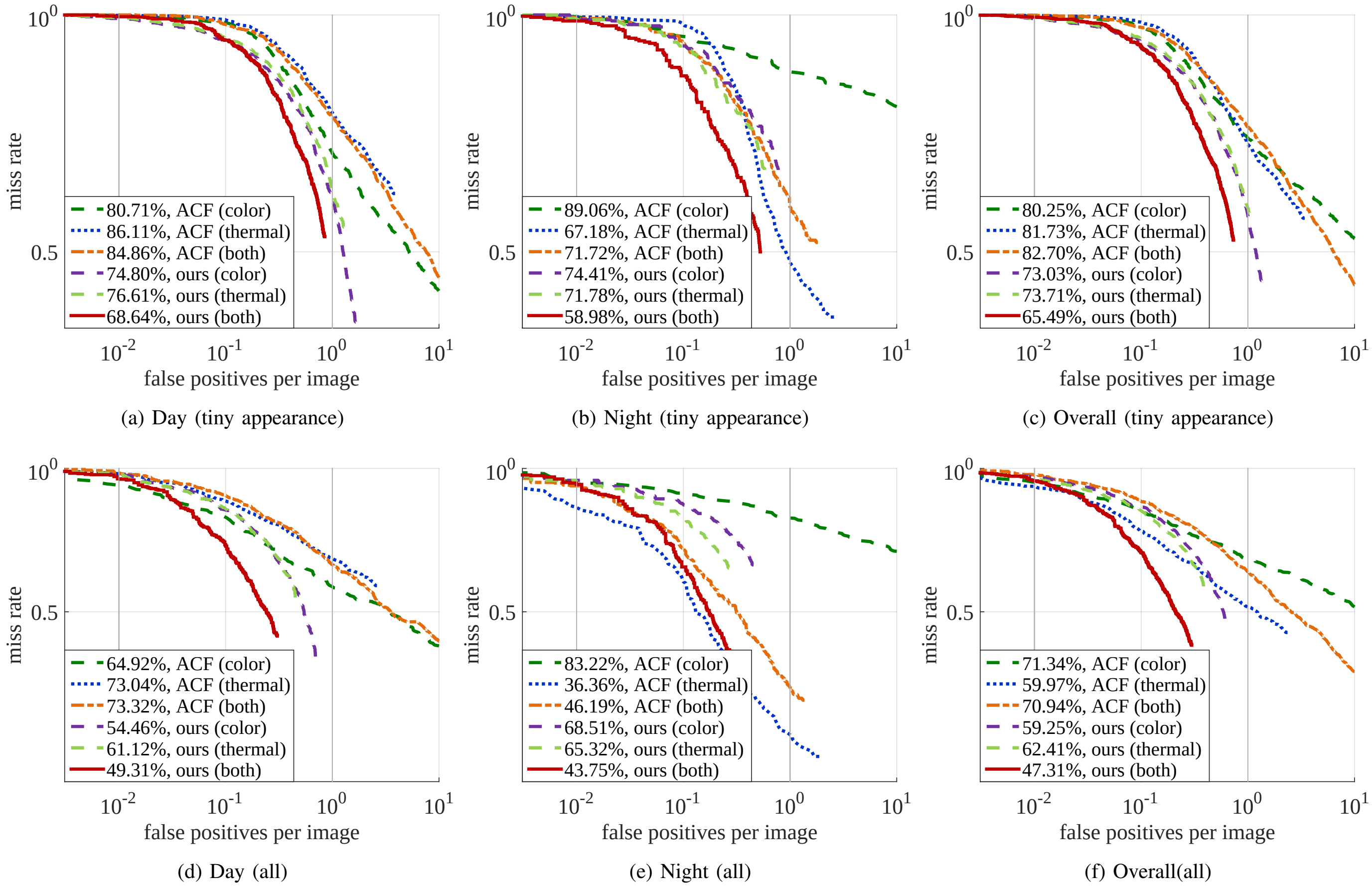
VI. Experimental Results and Discussion

A. Experimental settings

In our experiments, the proposed method was implemented with the following fixed parameter setings for all datasets: $\{\alpha, \beta, \gamma, \tau\} = \{1/3, 1/3, 1/3, 3/5\}$. We implemented the proposed models using the MatConvNet toolbox [11] and SVR [10]. Our multiple feature are extracted at the stage of 1,3, and 5. we employ KAIST pedestrian benchmark [8] for training and testing which provides aligned RGB and FIR images including day and night. For training, we did not use any normalization scheme both RGB and FIR images owing to batch normalization layer.

B. Evaluation on KAIST Multipsectral Benchmark

We select 20 frames from the KAIST Multispectral Benchmark [8] to remove redundancy. We divide datasets into training (2500 images) and testing (2252 images) set. We also disregard miss-aligned annotation images for calculate detection rate correctly. We strictly follow the evaluation



(a) Day (tiny appearance)

(b) Night (tiny appearance)

(c) Overall (tiny appearance)

(d) Day (all)

(e) Night (all)

(f) Overall(all)

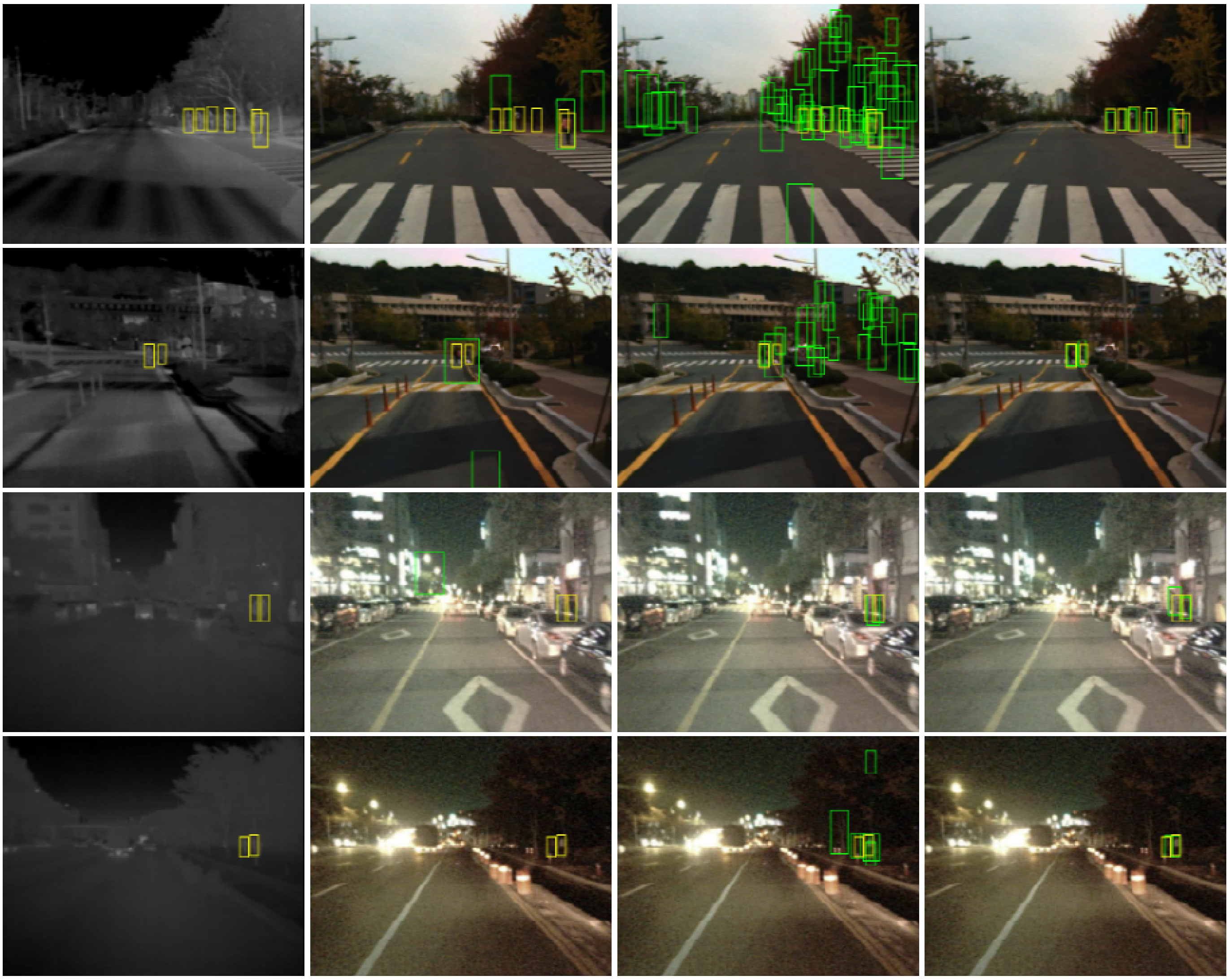
Fig. 5. Quantitative evaluations as false positive per image (FPPI) versus miss rate in various conditions.

protocol in [3], which measures the log average miss rate over 9 points ranging from $10^{-}2$ to 10^{0} false-positive-per-image.

Fig. 5 and Fig. 6 provide a quantitative and qualitative comparison of the proposed method to other state-of-the-art approaches, such as ACF [3] and DPM [5]. As shown in Fig. 5, the tiny appearance means that the size of pedestrian is below 30 width. Furthermore, thermal results were tested on only FIR domain. Note that we did not include DPM [5] on quantitative results because DPM [5] can not be trained for multiple channels. These hand-crafted frameworks mainly lack in discriminative power to cover different modality images, thus providing limited performance. Even more, hand-crafted frameworks also showed limitation on a small scale circumstances as in Fig. 1(c). Although ACF (thermal) is comparatively better results than ACF (both), ACF (thermal) are not stable on day results due to only use FIR images as in Fig. 5(d). Unlike these approaches, proposed method performed well and stable regardless the difference of day and night owing to the well-designed network, where we can obtain more discriminative features, and the joint usage of RGB-FIR images using SVR and accumulated map. We also performed a qualitative evaluation with results listed in Fig. 6, which also clearly demonstrates the effectiveness of proposed method. Specifically, similar to results in Fig. 5, ACF [3] and DPM [5] have a limitation to detect small scaled pedestrian, where the KAIST Multispectral Benchmark [8] includes small scaled pedestrians. Since DPM [5] are part-based method, they show unsatisfactory result. However, proposed method shows reasonable results on small scaled pedestrians. Moreover, proposed method, which effectively joint use of RGB-FIR images, we can detect pedestrian both day and night without any modifying algorithm.

VII. Conclusion

We introduced the method for pedestrian detection on both day and night. We employ CNN-based framework to have more discriminative power compared with conventional methods and the joint use of RGB-FIR images to combine these multi-spectral image into a single framework. To best of our knowledge, our approach is a novel and important task for pedestrian detection during not only the day, but also the night, which providing the reliability in many applications and the solution for troublesome of bad visibility at night effectively. Our approach also presented to localize the pedestrian using accumulated proposal-wise probability map. In the experimental results in comparison to state-of-the-art method dealing with multi-spectral images for pedestrian detection, we demonstrated our proposed method is more effective and robust pedestrian detection regardless of the day and night. In further work, the proposed method may be applied to various



(a) FIR image (b) DPM [5] (c) Multi-spectral ACF [8] (d) Ours

Fig. 6. Qualitative evaluation results at day and night. Estimated bounding box and ground-truth visualize green and yellow, respectively. Proposed method performs well even in a challenging circumstances at both day and night.

combinations of hard weather conditions, such as rain-and-night, cloudy and foggy at day.

REFERENCES

[1] M. Enzweiler and D.M. Gavrila. Monocular Pedestrian Detection: Survey and Experiments, In TPAMI, 31(12):2179-2195, 2009.

[2] N. Dalal and B. Triggs. Histograms of Oriented Gradients for Human Detection, In CVPR, 2005.

[3] P. Dollar, R. Appel, S. Belongie, and P. Perona. Fast feature pyramids for object detection, In TPAMI, 36(8):1532-1545, 2014.

[4] P.A. Viola and M.J. Jones, Robust Real-Time Face Detection. IJCV, 57(2):137-154, 2004.

[5] P.F. Felzenszwalb, R.B. Girshick, D. McAllester, and D. Ramanan. Object Detection with Discriminatively Trained Part Based Models, In TPAMI, 32(9):1627-1645, 2010.

[6] K. He, X. Zhang, S. Ren, and J. Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. In ECCV, 2014.

[7] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In CVPR, 2015.

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

[9] F. Xu, X. Liu and K. Fujimura. Pedestrian detection and tracking with night vision, In ITS, 6, 63-71, 2005.

[10] C. C. Chang and C. J. Lin. LIBSVM: a library for support vector machines, 2001.

[11] Online.: http://www.vlfeat.org/matconvnet/.

[12] S. Ioffe and C. szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In ICML, 2015.

[13] C. L. Zinick and P. Dollar. Edge Boxes: Locating object proposlas from edges. In ECCV, 2014.

[14] S. Kim, D. Min, B. Ham, S. Ryu, M. Do, and K. Sohn. DASC: Dense Adaptive Self-Correlation Descriptor for Multi-modal and Multi-spectral Correspondence, In CVPR 2015.

[15] Y. Tian, P. Luo, X. Wang, and X. Tang. Pedestrian detection aided by deep learning semantic tasks. In CVPR, 2015.

[16] S. Liao, X. Zhu, Z. Lei, L. Zhang and S. Z. Li. Learning Multi-scale Block Local Binary Patterns for Face Recognition, In PR, 2007.

[17] Y. Tian, P. Luo, X. Wang, and X. Tang. Pedestrian detection aided by deep learning semantic tasks. In CVPR 2015.