Journal of Imaging Science and Technology® 61(6): 060403-1–060403-9, 2017.

Pedestrian Detection at Night Using Deep Neural Networks and Saliency Maps

Duyoung Heo, Eunju Lee, and Byoung Chul Ko
Department of Computer Engineering, Keimyung University, Sindang-dong, Dalseo-gu, Daegu, 704-701, Korea
E-mail: niceko@kmu.ac.kr

Abstract. *This study focuses on real-time pedestrian detection using thermal images taken at night because a number of pedestrian–vehicle crashes occur from late at night to early dawn. However, the thermal energy between a pedestrian and the road differs depending on the season. We therefore propose the use of adaptive Boolean-map-based saliency (ABMS) to boost the pedestrian from the background based on the particular season. For pedestrian recognition, we use the convolutional neural network based pedestrian detection algorithm, you only look once (YOLO), which differs from conventional classifier-based methods. Unlike the original version, we combine YOLO with a saliency feature map constructed using ABMS as a hardwired kernel based on prior knowledge that a pedestrian has higher saliency than the background. The proposed algorithm was successfully applied to the thermal image dataset captured by moving vehicles, and its performance was shown to be better than that of other related state-of-the-art methods.* © *2017 Society for Imaging Science and Technology.*
[DOI: 10.2352/J.ImagingSci.Technol.2017.61.6.060403]

INTRODUCTION

In an advanced driver assistant system (ADAS), accurate pedestrian detection is the basic requirement for safe driving. Although many studies in this area have been conducted over the past few decades, most have focused on pedestrian detection during the day using color cameras. However, according to a report by the US National Highway Traffic Safety Administration in 2008,[1] the highest percentage of pedestrian fatalities occurs from 6 p.m. to 9 p.m., followed by 9 p.m. to midnight. In addition, the probability of a crash fatality is at its highest between 3 a.m. to 6 a.m. However, many studies related to pedestrian detection have been conducted during a daytime environment, and there is therefore a limit to applying daytime algorithms for a nighttime application.

For pedestrian detection at night, infrared (IR) based cameras, including near-IR (NIR) and far-IR (FIR) cameras, are commonly used instead of a normal charge-coupled device (CCD) camera because CCD camera is ineffective in environments with poor illumination, such as at nighttime. NIR camera is used in combination with an illuminator to see pedestrians at night, and is less expensive than an FIR camera. However, pedestrians become indistinguishable from the background similar to a CCD camera when pedestrian is located in front of vehicle headlights. Therefore, the present study uses an FIR camera because it is able to accurately discern the thermal energy emitted from a pedestrian's body regardless of the lighting conditions.[2]

In several studies, pedestrian detection in thermal images has been conducted using a hotspot[3,4] or region of interest (ROI).[5–7] Xu et al.[3] applied human detection by analyzing hotspot regions in a thermal image, including the face and shoulder areas, because these areas emit a large amount of thermal energy. However, because the basic assumption that the human body has a greater amount of thermal energy than the background is incorrect during the summer, and hotspot-based methods can only be used in a limited environment. ROI-based pedestrian detection methods use a similar assumption as hotspot-based methods in that the human body region is significantly brighter (or darker) than the background in a thermal image. To select the hotspots and ROIs, background subtraction,[5] adaptive thresholding,[3,4,7] and a gradient[6,8] are used. After extraction of the candidate hotspot and ROI regions, pattern classifiers such as a support vector machine (SVM) are applied to verify real pedestrians from other objects.

Recent studies have proposed the use of a saliency-model-based framework to segment pedestrians in thermal images because a saliency map provides crucial clues to discriminate important objects from the background. Numerous models and algorithms[3–5,9] for object detection have been proposed based on the assumption that salient objects in a color image tend to have a relatively high contrast in terms of color, luminance, and texture. Conventional saliency models measure the saliency based on measurements of the local center-surround (LCS)[10–15] and global contrast[12] cues. Although visual-contrast-based approaches are exemplified by their contrast capabilities with various types of handcrafted low-level features and tend to perform well in standard scenarios, they have a few limitations for challenging cases, for example, (1) the local contrast features may fail to detect homogeneous regions inside large salient objects, and (2) the global contrast suffers when the background is complex.[13]

Convolutional neural network (CNN) based approaches have recently been employed and have shown better saliency map results than visual-contrast-based methods.[14,15] However, all operations of these approaches are applied at the patch level which is a problem in typically blurry

Received June 25, 2017; accepted for publication Sept. 30, 2017; published online Nov. 29, 2017. Associate Editor: Zeev Zalevsky.
1062-3701/2017/61(6)/060403/9/$25.00

saliency maps without fine details, particularly near the boundary of a salient object. To resolve this problem, Li and Yu[13] proposed a different CNN-based approach consisting of two complementary components: a pixel-level fully convolutional stream, and a segment-wise spatial pooling stream. Although many CNN-based contrast networks provide improved saliency results, they commonly require longer time and more computational resources for the application of the two components. For more papers related to saliency map, please refer to.[16]

Although many studies related to saliency maps have been conducted for RGB images, there have been fewer studies on saliency maps used for thermal images. Yu et al.[17] used a bottom-up saliency algorithm for extracting feature maps at multiple spatial scales to segment areas of interest from an image. In this study, only two features maps are extracted based on Itti's local center-surround[10] method, i.e., a luminance map and an orientation map because a thermal image does not include color information. The SVM algorithm is then used to register a positive human saliency model for the trained classifiers. Ko et al.[4] used a luminance saliency map in thermal images to boost the boundary of the hotspots (head, hands, legs, etc.) by comparing the luminance contrast between the hotspots and background. The texture feature is then extracted from the luminance saliency map of a hotspot and applied to a random forest classifier for pedestrian detection. However, the thermal temperature between a pedestrian and the road is significantly different depending on the season. For example, during the summer, a pedestrian's thermal energy is almost similar to that of the road. Therefore, a saliency map may contain many false responses if the algorithm uses only one model without considering the season.

For pedestrian detection, conventional pattern classifiers such as an SVM[3] and a random forest[4] are popular algorithms. However, a CNN-based pedestrian detector[18] has recently achieved substantially better results than conventional classification algorithms.

Contributions of this Work

In this paper, we focus on pedestrian detection using a saliency map and CNN model for only thermal images. First, instead of a local center-surround-based approach, we adopt a Boolean-map-based saliency (BMS) approach[12] derived from the Boolean-map theory of visual attention to generate a global saliency map. However, the original BMA algorithm may be inappropriate for use in changeable thermal images depending on the season. To generate an optimal saliency map regardless of the season, we propose an adaptive BMS (ABMS) to distinguish the season and apply two different BMS models accordingly. Second, we use the CNN-based pedestrian detection algorithm, which differs from conventional classifier-based methods. In this study, you only look once (YOLOv2)[19] algorithm, which is a real-time object detection system used to recognize various pedestrians in a real-life road environment, is applied. Unlike the original YOLOv2, we combine YOLOv2 with a saliency feature map constructed using the proposed ABMA for training and testing purposes. ABMA is applied to the input image as a hardwired kernel based on prior knowledge that a pedestrian has a higher saliency than the background. The proposed method was successfully applied to the Keimyung University Pedestrian Detection (KMU-PD) dataset,[9] which was captured from moving vehicles for pedestrian detection, and its accuracy was confirmed to be higher than that of other related methods.

The remainder of this paper is organized as follows. Adaptive Boolean-map-based saliency detection describes the preprocessing for ABMS during different seasons. 'YOLOv_2 with ABMS for pedestrian classification' presents a pedestrian detection algorithm using YOLOv2 with ABMS. The differences in the experimental results between the proposed algorithm and other related state-of-the-art methods are described in Experimental Results. Finally, Conclusion and future works provides some concluding remarks regarding this research, along with areas of future work.

ADAPTIVE BOOLEAN-MAP-BASED SALIENCY DETECTION

Contrast-based saliency detection is used to detect conspicuous regions in multi-scale feature maps, followed by normalization and fusion of the resulting conspicuity maps.[20] However, feature maps are not suitable for measuring the surroundedness because they tend to lose the topological structure of the scene by only sparsely highlighting certain local patterns (e.g., edges or corners), and have a higher saliency for a complex background than for a salient object. Therefore, we use BMS,[12] which is less responsive to the elements in the background using global structural information.

Given image I, the BMS algorithm assumes that a set of Boolean maps $\mathbf{B} = \{B_1, B_2, \ldots, B_N\}$ in BMS are generated from distribution function $F(\mathbf{B}|I)$ conditioned on input image I, and that a set of attention maps $\mathbf{A}(\mathbf{B}) = \{A(B)_1, A(B)_2, \ldots, A(B)_N\}$ represent the influence of Boolean map Bi on the visual attention. The saliency is then modeled using the mean attention map $\bar{A}$ over randomly generated Boolean maps using Eq. (1).

$$\bar{A} = \int_{i=1}^{N} A(B)_i \, \mathrm{d}F(B_i|I). \tag{1}$$

Boolean maps $\mathbf{B}$ are constructed by randomly thresholding the input's feature map $\emptyset(I)$ using the $\mathbf{THRESH}(\cdot, \theta)$ function, according to the prior distributions over the feature channel and threshold θ.

$$B_i = \mathbf{THRESH}(\emptyset(I), \theta) \tag{2}$$

where the $\mathbf{THRESH}(\cdot, \theta)$ function assigns 1 to a pixel if its value on the input map is greater than θ, and zero otherwise. In 12, feature channels can consist of multiple features such as the color, intensity, depth, and motion. To generate Boolean maps, the study in 12 enumerates a few color channels, and

samples threshold θ from zero to 255 with a fixed step size of δ. An inverted copy of each Boolean map is also included in the output to account for the inverted region selection. The opening operation is then applied to each Boolean map for noise removal.

However, the pedestrians in thermal images have a different saliency according to the season, and we therefore modified the method for constructing a Boolean map as follows:

1. Because this study uses thermal images, the feature channel consists solely of a luminance channel.
2. Because the initial threshold and sampling step should be changeable according to the emitting energy of the road area, we determine the prior distribution function for the initial threshold t_1 and sampling step Δs of the luminance channel using Eqs. (3) and (4).

$$t_1 = \max(C, \mu + \sigma) \tag{3}$$

where C,μ, and σ represent the minimum constant value of the pedestrian, and mean and standard deviations of the road, respectively. Inspired by,[2] we first establish the front part of the car as the reference region as shown in Figure 1(a); the instantaneous stimulus μ and σ is estimated by averaging mean and standard variation of the thermal intensity during five frames. The sampling step Δs is determined based on t_1 and the size of Boolean map N:

$$\Delta s = \frac{255 - t_1}{N} \tag{4}$$

where N is the number of attention maps and it is an adjustable parameter according to the system requirements; the accuracy and processing time of the pedestrian detection largely depend on the value of N. For example, the processing time is increased steadily as the value of N is increased. In contrast, the accuracy decreases as the value of N is decreased. In this study, we set N to 5 according to the experiment results. The detailed experiment conducted to select an optimal N is described in Experimental Results.

From Eqs. (3) and (4), initial threshold t_1 and sampling step Δs can have different values according to the season. For example, it is reasonable to start the threshold at a higher value with a denser sampling step during the summer than during the winter because the road emission energy is high and has a similar temperature with the pedestrians during the summer.

After detecting the initial threshold and sampling step, all N attention maps are normalized to obtain a greater emphasis on small concentrated active areas before the linear combination step. All attention maps are linearly combined into full-resolution mean attention map $\bar{A}$. After normalization of the mean attention map, post-processing is applied to $\bar{A}$ to produce saliency map S through Gaussian blurring.

The algorithm used to construct a saliency map by applying the proposed ABMS is described in Algorithm 1 and illustrated in Figure 1.

Algorithm 1: Adaptive BMS

A set of Boolean maps $\mathbf{B} = \{\}$; A set of attention maps $\mathbf{A} = \{\}$; $\bar{A} \leftarrow 0$

(1) Input feature map $\emptyset(I)$ is generated from a thermal image

Compute the initial threshold t_1 and sampling step Δs using Equations (3) and (4)

For i = 1 to N // refer to the results in Figure 1(b), (c)

For $\theta = t_1$ to 255

$B_i = \mathbf{THRESH}(\emptyset(I), \theta)$

$\bar{B}_i = \mathbf{INVERT}(B_i)$

Morphological opening to B_i and $\bar{B}_i$

Add B_i and $\bar{B}_i$ to $\mathbf{B}$

End for

End for

(2) For i = 1 to N // refer to the results in Figure 1(d)

Set $A_i(x, y) = 0$ if all pixels of $B_i(x, y)$ are connected to the image borders

Morphological dilation to A_i

Normalization A_i

$\bar{A} \leftarrow \bar{A} + A_i$

End for

(3) $\bar{A} \leftarrow \bar{A}/\max_i A_i$ // refer to the results in Figure 1(e)

(4) S$\leftarrow$ $PostProcessin(\bar{A})$ // refer to the results in Figure 1(f)

Fig. 1 shows the overall ABMS procedure for two different seasons. Based on the results of each season, we can determine that a pedestrian has high saliency compared to the neighborhood regions regardless of the season.

YOLOv2 WITH ABMS FOR PEDESTRIAN CLASSIFICATION

In previous studies on pedestrian detection, many classifier-based methods[3–7] have generally proposed to apply a single trained classifier or a combination of several classifiers to multiple locations as a sliding window method or scale sliding window method within a single image. Thus, only certain sliding windows with high scoring are selected. A CNN, a type of deep learning method, has recently become a very popular classification algorithm in computer vision, artificial intelligence, and pattern recognition fields. A CNN is a multi-layer feed forward neural network, and can learn multiple stages of invariant features using a combination of supervised and unsupervised learning methods. In particular, LeCun et al.[21] have developed a multi-layer CNN for real-world situations, and have demonstrated excellent performance of hand-written digit classification. Since the successful results of,[21] the use of a CNN has shown impressive progress in computer vision research, particularly object and pedestrian detection. Although CNN-based object detection has shown better results than conventional classifier-based approaches, the computational expense is a major problem. To reduce the computational cost of a CNN-based approach while maintaining the detection performance, Fast R-CNN[24] and

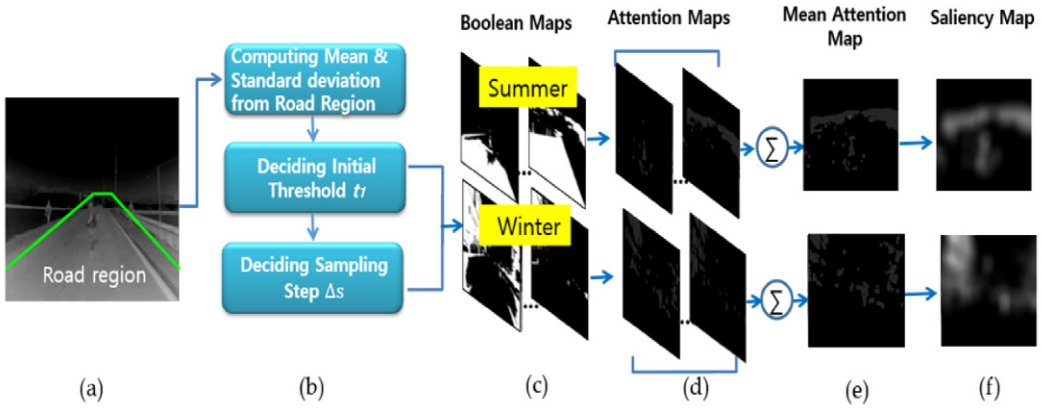

Figure 1. The overall ABMS procedure: (a) input thermal image, (b) adaptive determination of the initial threshold and sampling step according to the thermal energy of the road region, (c) Boolean maps for two seasons, (d) attention maps for two seasons, (e) mean attention maps for two seasons, and (f) final saliency maps.

Figure 2. Illustration of tiny YOLOv2 adopting a handcrafted ABMS kernel. Tiny YOLOv2 consists of nine convolutional layers and six max pooling layers followed by two fully connected layers. In our system, ABMS is applied as the initial handcrafted kernel. The input image is first applied to the ABMS kernel, and a saliency map is transferred to the YOLO network.

Faster R-CNN[22] have been proposed, which depend on a region-based CNN and the sharing of convolutions to hypothesize the object locations.

A real-time object detection system based on a single convolutional network, i.e., YOLOv1,[23] was recently proposed. YOLOv1 uses a single neural network to predict the bounding boxes and class probabilities directly from full images in a single evaluation. Because the entire detection pipeline is a single network, the detection performance end-to-end can be quickly and directly optimized. A YOLOv1 network consists of 24 convolutional layers followed by two fully connected layers, and uses 1×1 reduction layers followed by 3×3 convolutional layers. YOLOv1 also has a fast version, i.e., tiny YOLOv1, which was designed to push the boundaries of fast object detection. This fast YOLOv1 uses a neural network with fewer convolutional layers (nine instead of 24) and fewer filters in these layers. Although YOLOv1 has fast detection time, it has a relatively low recall rate compared to region-based methods (Fast[24] and Faster R-CNN[22]). To improve the detection accuracy of YOLOv1, a new version of YOLOv2[18] was proposed. YOLOv2 is basically constructed based on YOLOv1, including the layer structure, 1×1 reduction layers, and 3×3 convolutional layers. However, it applies additional ideas to YOLOv1, such as anchor boxes, dimension clusters, batch normalization, and a high-resolution classifier, to improve the accuracy and reduce the speed. In this study, we use a fast version of YOLOv2, i.e., tiny YOLOv2, instead of the original version, to detect pedestrians in real-time applications.

As Figure 2 shows, we use the tiny YOLOv2 architecture with nine convolutional layers and six max pooling layers followed by two fully connected layers. The difference from the original tiny YOLOv2 is that our approach adopts ABMS as the handcrafted kernel to emphasize the pedestrians from the background. The modified tiny YOLOv2 predicts the detection in a 13×13 feature map, and predicts five bounding boxes at each cell in the output feature map.

EXPERIMENTAL RESULTS

Despite a large number of datasets for pedestrian detection in color video sequences,[25–28] only a small number of thermal test datasets[9,29] have been introduced. In this study, we use the KMU-PD dataset,[9] which was captured using a thermal camera from moving vehicles (at 20 to 30 km/h) during the summer and winter for pedestrian detection. KMU-PD contains images involving a moving camera, moving pedestrians, sudden deformations in shape, unexpected changes in motion, and partial or full occlusions

Table I. Comparison of computational speeds for the six methods.

Methods	(1) HOG+SVM[30]	(2) Hotspot +SVM[3]	(3) CS-LBP+Hotspot+RF[4]	(4) OCS-LBP+RF[9]	(5) T-YOLOv2[19]	(6) Proposed method
FPS	10.7	11.2	9.8	16.56	63.8	62.8

between pedestrians at night, with varying pedestrian speeds and activities.

The positive training data consisted of 3,795 thermal images including pedestrians covering a wide variety of sizes and poses. The negative data consisted of 3,148 thermal images including those randomly cropped from the background. To evaluate the detection performance of the proposed system, we used eight video sequences of KMU-PD. The total amount of test data for winter (video sequences 1 through 4) is 1,069 frames, and the total amount for summer (video sequences 5 through 8) is 1,400 frames. Each video sequence contains various pedestrian situations according to the particular scenario, such as crossing the road or walking down the sidewalk. For training tiny YOLOv2, we used a pre-trained convolutional weights that are trained on ImageNet as the initial parameters of the system. Then, we performed fine tuning of the network based on KMU-PD. In the fine tuning process, we trained the network with learning rate 10^{-2} for 965 epochs, a batch size of 64 and momentum of 0.9.

All experiments were conducted on an Intel Core i7 PC with 8 GB of RAM running Microsoft Windows 7. For the graphical processing unit (GPU), we used an NVIDIA GTX 970. All experiments were conducted based on the GPU and Compute Unified Device Architecture (CUDA) for deep learning.

Comparison of Pedestrian Detection Methods

To evaluate the detection performance, we used precision and recall measures that are generally applied to evaluate the human detection performance. The detection results are assigned to the ground-truth objects and verified as true/false positives or negatives by measuring the bounding box overlap. As the detection criteria, the detected bounding box is considered to be a correct detection result if the overlap ratio between the detected bounding box and the ground-truth bounding box exceeds 50%.

We compared the performance of the proposed pedestrian detection method with that of five other state-of-the-art methods: (1) HOG features with a linear SVM[29] (HOG+SVM), (2) hotspot with intensity features and SVM[3] (Hotspot + SVM), (3) hotspot with CS-LBP features with a random forest[4] (CS-LBP+Hotspot+RF), (4) OCS-LBP with a random forest[9] (OCS-LBP+RF), (5) tiny YOLOv2[19] (T-YOLOv2), and (6) ABMS with tiny YOLOv2 (the proposed method).

Figure 3 shows the results of the performance comparison between these six approaches. According to the result, HOG+SVM[30] achieved the worst detection results

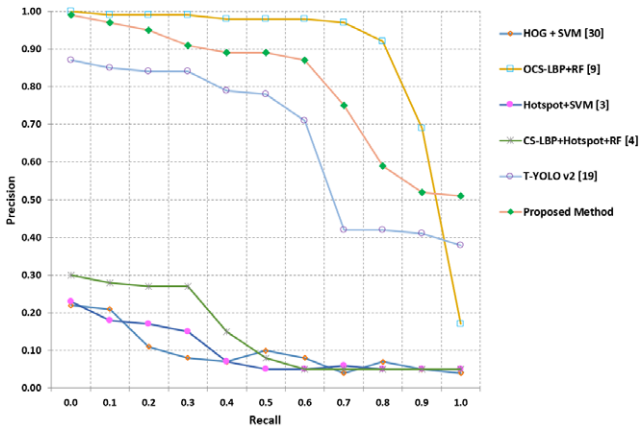

Figure 3. Performance comparison of precision versus recall using the same KMU-PD dataset.

in terms of precision and recall because, unlike in a color image, HOG cannot characterize pedestrians in a thermal image. Hotspot-based methods, i.e., Hotspot + SVM[3] and CS-LBP+Hotspot+RF[4] achieved a somewhat better detection performance than HOG+SVM. However, these approaches still have difficulty in detecting pedestrians during the summer because the hotspot regions cannot be detected. The OCS-LBP+RF approach[9] showed a better performance than T-YOLOv2-based methods because it used a number of decision trees and bootstrapping steps for improving performance. However, the slow processing time is a major limitation of this approach compared to T-YOLOv2-based approaches because its computational speed was around 46 fps slower than that of YOLOv2-based approaches as shown in Table I.

The proposed approach produced the second highest precision and recall rate as 87.12% and 62%, which are an 8.64% and 12% higher precision and recall than T-YOLOv2. Based on these results, we confirmed that ABMS is a reasonable preprocessing approach to characterizing a pedestrian's body from the background. In addition, we knew that T-YOLOv2 with ABMS is fast and an accurate pedestrian detector, although they use shorter layers instead of the full layers of YOLOv2.

To achieve a fair performance comparison of the same six algorithms by using an additional dataset, we also performed pedestrian detection using the state-of-the-art CVC-09 FIR dataset.[29] CVC-09 FIR dataset is composed by two sets of images, named as the day and night sets, which refers to the moment of the day they were acquired. This dataset contains 5,308 frames for training and 5,763

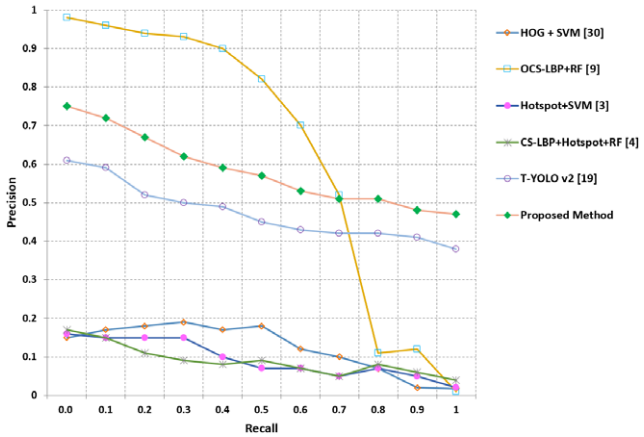

Figure 4. Performance comparison of precision versus recall using the same CVC-09 dataset.

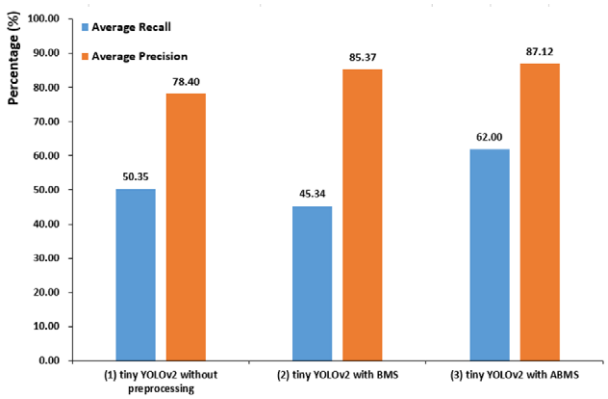

Figure 5. Comparison of average precision for three different approaches.

frames for testing. For training proposed approach, we used the same condition with KMU-PD dataset; we performed a fine tuning of the network based on CVC-09 FIR dataset with learning rate 10^{-2} for 965 epochs, a batch size of 64, and a momentum of 0.9.

Figure 4 shows the precision and recall rate for the six methods. OCS-LBP+RF[9] showed highest precision and recall rate (69%, 53%) and the proposed method produced the second highest precision and recall rate (59%, 39%) as similar pattern to those shown in Fig. 3. The original T-YOLOv2[19] produced the third highest precision and recall rate as 50% and 26%, which are a 9% and 13% lower precision and recall than the proposed method. Although algorithm OCS-LBP+RF[9] showed higher precision results than those of the proposed method, precision dropped rapidly when the recall value was 0.8. HOG+SVM[30] and Hotspot + SVM[3] also achieved the worst detection results in terms of precision and recall rate similar to the results of the first experiment. From Fig. 4, we knew that the precision and recall rate in all algorithms is commonly low compared to the result of Fig. 3 because CVC-09 FIR dataset consists of day and nighttime video sequences and has more confuse background objects (road, building, cars, etc.) representing similar thermal energy with pedestrians than KMU-PD dataset. By conducting experiments using the additional CVC-09 FIR dataset, we verified that the proposed method was not over-tuned for the KMU dataset and gave a generalized performance for various environments by applying ABMS as a preprocessing. However, the low recall rate in daytime of CVC-09 FIR dataset is a main problem to be solved for designing more reliable system in the future work.

Based on these results, we confirmed that ABMS is a reasonable preprocessing step for a YOLO architecture to differentiating a pedestrian's body from the background.

In addition to the precision and recall, the computational speed of the proposed method was compared with that of the five comparison methods using KMU-PD dataset. As shown in Table I, the computational speed of the proposed method was approximately 52.1, 51.6, 53, and 20.4 frames faster than that of the classifier-based methods, i.e., HOG+SVM,[30] Hotspot+SVM,[3] CS-LBP+Hotspot+RF,[4] and OCS-LBP+RF,[9] respectively, using the same test images. The T-YOLOv2 method achieved similar or slightly faster computational time (63.8 fps) compared to the proposed method (62.8 fps) because the proposed method requires more time for ABMS processing. Based on these results, we confirm that the YOLOv2-based approaches, which consist of a number of convolutional layers, have fast processing time because they use the GPU along with the CUDA technique for convolution processing.

The OCS-LBP+RF[9] method achieved a similarly high detection performance as the proposed approach, as shown in Fig. 3. However, because its computational speed was still around 46 fps slower than that of YOLOv2-based methods, it is less suitable for real-time systems as compared to the proposed method.

Performance Evaluation of ABMS

To prove the effectiveness of ABMS, we examined its average precision for the following three cases using KMU-PD dataset: (1) tiny YOLOv2 trained using KMU-PD without a preprocessing step, (2) tiny YOLOv2 trained using KMU-PD with BMS preprocessing, and (3) tiny YOLOv2 trained using KMU-PD with ABMS preprocessing.[1]

As shown in Figure 5, we confirm that the proposed approach shows 11.6% and 16.6% higher average recall rate than cases (1) and (2). In terms of the average precision, the proposed approach also shows 8.72%, and 1.75% higher rate than cases (1) and (2), respectively. The proposed approach shows a good detection performance because, instead of a raw image, it uses handcrafted kernel ABMS as a preprocessing step to characterize the contrast of the pedestrian from the background. In addition, ABMS showed

[1] Because the accuracy of the BMS algorithm has already been demonstrated in previous papers,[12] we only compared the performance on three cases.

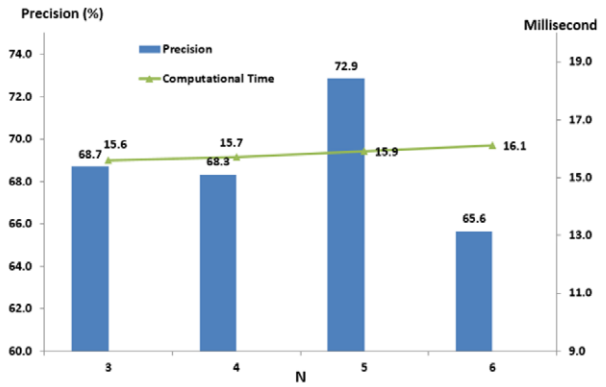

Figure 6. Four possible values for determining the optimal number of Boolean maps N in terms of precision and computational time.

a better performance than BMS because it uses different threshold and sampling steps when considering the season.

Determination of Boolean-Map Size

For ABMS, the size of Boolean map N is an important factor for improving the detection performance because the sampling step is determined based on this size. Here, N is an adjustable parameter according to the system requirements, such as the speed and accuracy. For example, if the value of N is too large, the processing time increases linearly. In contrast, if the value of N is too small, the method may not converge to the target location. Therefore, it is essential to determine the value of N for improving both the accuracy and processing speed. Therefore, we determine the appropriate value of N by estimating the average detection precision and computational time using four candidate values of N.

As shown in Figure 6, when N is set to 5, the precision is the highest when compared with the other values. In particular, when N is less than 5, the precision is also decreased because the contrast between the pedestrian and background is not large. In contrast, when N is larger than 5, the degradation in performance is increased more than at smaller values. However, the computational time of four candidate cases are almost similar. The number of Boolean maps is closely related to the accuracy, which we set to 5 according to the results shown in Fig. 6.

Figure 7 shows the pedestrian detection results obtained using our proposed method on the KMU-PD dataset. The detection results show that our proposed method detects each pedestrian correctly, regardless of the season, cluttered background, distance from the camera, or pedestrian overlap. However, false detections caused by a hot car region or a hot background region (e.g., lamp or electric pole) should be resolved in the next version.

CONCLUSION AND FUTURE WORKS

This paper proposed a pedestrian detection algorithm using a CNN model based on YOLOv2 and saliency maps for thermal images. Because pedestrians emit different amounts of energy depending on the season, we applied two different BMS models and saliency maps adaptively according to the season as a preprocessing of the CNN model. For the pedestrian detection, we used the YOLOv2 algorithm, which is a real-time object detection system used to recognize various pedestrians in a real-life road environment. Unlike the original YOLOv2, we combine YOLOv2 with a saliency feature map constructed using the proposed ABMA for training and testing purposes. ABMA is applied to the input image as a hardwired kernel based on prior

Figure 7. Sample pedestrian detection results using the proposed method during (a) summer, (b) winter, and (c) false detection in KMU-PD dataset

knowledge that a pedestrian has higher saliency than the background.

The experimental results, when using the KMU-PD dataset, show that the proposed method can detect pedestrians correctly in real time regardless of the season. However, the proposed method still incurs some false positives or missing detections in the CVC-09 dataset when the thermal energy of the pedestrian is the same as that of the road region, when a severe overlap with other pedestrians occurs, and for the presence of pedestrians in daytime with a faint appearance.

Apart from detecting a pedestrian, semantic segmentation, which is understanding an image at pixel level based on deep learning[31–33] has been researching for semantic object segmentation. Although the results of semantic segmentation are more detail than object detection, it is effective only in a simple image and needs more computational time than normal CNN because of convolution and extra deconvolution. Therefore, we plan to replace the ABMA into semantic segmentation by modifying[33] to produce coarse segmentation maps and to predict the rough semantic location of a pedestrian in a short time.

Moreover, we plan to optimize tiny YOLOv2 without degrading the accuracy or speed to be suitable for small and affordable computers, such as Raspberry Pi, for real vehicle applications.

ACKNOWLEDGMENT

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2016R1D1A1A09916581) and partially supported by Future Vehicle Leading Technology Development Project funded by Daegu City (DG-2017-01).

REFERENCES

[1] National Highway Traffic Safety Administration. National pedestrian crash report. National Technical Information Service (2008).

[2] J. Y. Kwak, B. C. Ko, and J. Y. Nam, "Pedestrian tracking using online boosted random ferns learning in far infrared imagery for safe driving at night," IEEE Trans. Intell. Trans. Syst. **18**, 69–81 (2017).

[3] F. Xu, X. Liu, and K. Fujimura, "Pedestrian detection and tracking with night vision," IEEE Trans. Intell. Trans. Syst. **6**, 63–71 (2005).

[4] B. C. Ko, D. Y. Kim, and J. Y. Nam, "Detecting human using luminance saliency in thermal images," Opt. Lett. **37**, 4350–4352 (2012).

[5] E. S. Jeon, J. S. Choi, J. H. Lee, K. Y. Shin, Y. G. Kim, T. T. Le, and K. R. Park, "Human detection based on the generation of a background image by using a far-infrared light camera," Sensors **15**, 6763–6788 (2015).

[6] Y. Ma, X. Wu, G. Yu, Y. Xu, and Y. Wang, "Pedestrian detection and tracking from low-resolution unmanned aerial vehicle thermal imagery," Sensors **16**, 446–472 (2016).

[7] J. Ge, Y. Luo, and G. Tei, "Real-time pedestrian detection and tracking at nighttime for driver-assistance systems," IEEE Trans. Intell. Trans. Syst. **10**, 283–298 (2009).

[8] X. Y. Zhao, Z. X. He, S. Y. Zhang, and D. Liang, "Robust pedestrian detection in thermal infrared imagery using a shape distribution histogram feature and modified sparse representation classification," Pattern Recognit. **48**, 1947–1960 (2015).

[9] M. Jeong, B. C. Ko, and J. Y. Nam, "Early detection of sudden pedestrian crossing for safe driving during summer nights," IEEE Trans. Circuits. Syst. Video Technol. **27**, 1368–1380 (2017).

[10] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," IEEE Trans. Pattern Anal. Mach. Intell. **20**, 1254–1259 (1998).

[11] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," *Proc. Advances in Neural Information Processing Systems–NIPS 2007* (NIPS Foundation Inc., Vancouver, Canada, 2007), pp. 545–552.

[12] J. Zhang and S. Sclaroff, "Exploiting surroundedness for saliency detection: a Boolean map approach," IEEE Trans. Pattern Anal. Mach. Intell. **38**, 889–902 (2016).

[13] G. Li and Y. Yu, "Deep contrast learning for salient object detection," *Proc. 2016 IEEE Int'l. Conf. on Computer Vision and Pattern Recognition–CVPR 2016* (IEEE, Piscataway, NJ, 2016), pp. 478–487.

[14] R. Zhao, W. Ouyang, H. Li, and X. Wang, "Saliency detection by multi-context deep learning," *Proc. 2016 IEEE Int'l. Conf. on Computer Vision and Pattern Recognition–CVPR 2015* (IEEE, Piscataway, NJ, 2015), pp. 1265–1274.

[15] L. Wang, H. Lu, X. Ruan, and M. H. Yang, "Deep networks for saliency detection via local estimation and global search," *Proc. 2015 IEEE Conf. on Computer Vision and Pattern Recognition–CVPR 2015* (IEEE, Piscataway, NJ, 2015), pp. 3183–3192.

[16] Q. Zhao and C. Koch, "Learning saliency-based visual attention: a review," Signal Process. **93**, 1401–1407 (2013).

[17] D. Yu, J. Han, Y. Ye, and Z. Fang, "A novel saliency detection framework for infrared thermal images," *Proc. 2014 IEEE Int'l. Conf. Orange Technologies–ICOT 2014* (IEEE, Piscataway, NJ, 2014), pp. 57–60.

[18] J. Liu, S. Zhang, S. Wang, and D. N. Metaxas, "Multispectral deep neural networks for pedestrian detection," *Proc. 2016. British Machine Vision Conference–BMVC 2016* (BMVA Press, York, 2016), pp. 1–13.

[19] J. Redmon and A. Farhadi, "YOLO9000: better, faster, stronger," *Proc. 2017 IEEE Int'l. Conf. on Computer Vision and Pattern Recognition-CVPR 2017* (IEEE, Piscataway, NJ, 2017), pp. 1–9.

[20] J. Zhang and S. Sclaroff, "Saliency detection: a Boolean map approach," *Proc. 2013 IEEE Int'l. Conf. on Computer Vision-ICCV 2013* (IEEE, Piscataway, NJ, 2013), pp. 153–160.

[21] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," Neural Comput. **1**, 541–551 (1989).

[22] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," IEEE Trans. Pattern Anal. Mach. Intell. **99**, 1–14 (2016).

[23] J. Redmon, S. Divvala, S. Girshick, and A. Farhadi, "You only look once: unified, real-time object detection," *Proc. 2016 IEEE Int'l. Conf. on Computer Vision and Pattern Recognition–CVPR 2016* (IEEE, Piscataway, NJ, 2016), pp. 1–10.

[24] R. Girshick, "Fast R-CNN," *Proc. 2015 IEEE Int'l. Conf. on Computer Vision–ICCV 2015* (IEEE, Piscataway, NJ, 2015), pp. 1–9.

[25] A. Ess, B. Leibe, and L. V. Gool, "Depth and appearance for mobile scene analysis," *Proc. 2007 IEEE Int'l. Conf. on Computer Vision–ICCV 2007* (IEEE, Piscataway, NJ, 2007), pp. 1–8.

[26] C. Wojek, S. Walk, and B. Schiele, "Multi-cue onboard pedestrian detection," *Proc. 2009 IEEE Int'l. Conf. on Computer Vision and Pattern Recognition–CVPR 2009* (IEEE, Piscataway, NJ, 2009), pp. 794–801.

[27] M. Enzweiler and D. M. Gavrilla, "Monocular pedestrian detection: survey and experiments," IEEE Trans. Pattern Anal. Mach. Intell. **31**, 2179–2195 (2009).

[28] P. Dollár, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: an evaluation of the state of the art," IEEE Trans. Pattern Anal. Mach. Intell. **34**, 743–761 (2012).

[29] Y. Socarrás, S. Ramos, D. Vázquez, A. M. López, and T. Gevers, "Adapting pedestrian detection from synthetic to far infrared images," *Proc. 2013 ICCV Workshop on Visual Domain Adaptation and Dataset Bias–WDADB 2013* (IEEE, Piscataway, NJ, 2013), pp. 1–3.

[30] Y. Xu, D. Xu, S. Lin, T. X. Han, X. Cao, and X. Li, "Detection of sudden pedestrian crossings for driving assistance systems," IEEE Trans. Syst. Man Cyber. B **42**, 729–739 (2012).
[31] L-C. Chem, Y. Yang, J. Wang, W. Xu, and A. L. Yuille, "Attention to scale: scale-aware semantic image segmentation," *Proc. 2016 IEEE Int'l. Conf. on Computer Vision and Pattern Recognition–CVPR 2016* (IEEE, Piscataway, NJ, 2016), pp. 3640–3649.
[32] H. Zhao, J. Shi, Z. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," *Proc. 2017 IEEE Int'l. Conf. on Computer Vision and Pattern Recognition–CVPR 2017* (IEEE, Piscataway, NJ, 2017), pp. 1–11.
[33] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," IEEE Trans. Pattern Anal. Mach. Intell. **39**, 640–651 (2017).