International Journal of Electrical and Computer Engineering (IJECE)
Vol. 14, No. 4, August 2024, pp. 3962~3970
ISSN: 2088-8708, DOI: 10.11591/ijece.v14i4.pp3962-3970

Comparing Mask R-CNN backbone architectures for human detection using thermal imaging

Tan Dat Trinh, Pham Cung Le Thien Vu, Pham The Bao
Department of Computer Science, Information Science Faculty, Sai Gon University, Ho Chi Minh City, Vietnam

Article Info

Article history:

Received Oct 15, 2023
Revised Apr 1, 2024
Accepted Apr 16, 2024

Keywords:

Convolutional neural network
Deep learning
Faster R-CNN
Human detection
Mask R-CNN
Thermal image

ABSTRACT

We introduce a method for detecting humans in thermal imaging using an end-to-end deep learning model. Our objective is to optimize the human detection process in thermal imaging by investigating the mask region-based convolutional neural network (Mask R-CNN). The model, an advancement of the faster region-based convolutional neural network (Faster R-CNN), not only captures bounding boxes encompassing human subjects but also delineates segmentation masks around them. Our investigation extends to the evaluation and comparison of various convolutional neural networks for feature learning, like residual network (ResNet) and Inception ResNet, all integrated into the Mask R-CNN framework. Furthermore, the experimental results show that our proposed technique achieves high accuracy. Specifically, the Mask R-CNN model using ResNet50-V1 achieved the best results, with an F-value of 87.85%, a recall of 79.33%, and a precision of 98.41%.

Corresponding Author:

Pham The Bao
Department of Computer Science, Information Science Faculty, Sai Gon University
273 An Duong Vuong Street, Ward 3, District 5, Ho Chi Minh City, Vietnam
Email: ptbao@sgu.edu.vn

1. INTRODUCTION

The demand for human/pedestrian detection based on computer vision has risen significantly in various domains, including security, intelligent monitoring, search and recuse, surveillance systems, and other fields. Consequently, human/pedestrian detection has become a challenging and important field, attracting significant interest from the research community in recent years [1]. The most challenging task is the detection and identification of human activity during continuous 24/7 operations in environments such as border controls, school campuses, stations, and airports [2]. A visual camera can be set up during daytime operations, but it has obvious limitations in nighttime or dark environments. On the other hand, the visual camera is very sensitive to illumination effects. Thermal cameras (or infrared cameras) can help solve lighting problems, and thermal imaging with infrared cameras can enhance human visibility in low-light, dark or obscured environments.

Thermal imaging systems are utilized in real-time applications across military, industrial, and commercial sectors. This technology enhances surveillance capabilities, allowing for the detection of humans in low-light conditions, which is important for border security and urban safety. Additionally, in search and rescue missions, it efficiently identifies people in difficult environments, ultimately aiding in life-saving efforts. In the security and surveillance, it assists in monitoring people in crowded areas such as school campuses, stations, and airports. It helps identify potential threats and manage public events safely. Specifically, the detection of pedestrians in school campus surveillance is essential for safeguarding students, staff, and visitors. This system can enhance situational awareness, enabling rapid response to emergencies.

Therefore, thermal cameras are expected to work all the time. However, thermal image processing remains a challenging task due to several factors, including lower signal-to-noise ratio (SNR), polar inversion, reflection, and halo effects [3], [4]. Traditional human detection systems were built on handcrafted features and rely on machine learning algorithms. Their performance is often susceptible to the requirement of constructing intricate sets comprising numerous low-level features combined with high-level context from detectors and context classifications. With the development of deep learning, more powerful tools that can learn higher-level and deeper semantic features are being introduced to solve problems that exist in traditional architectures [3].

In traditional approaches, human detection in thermal imaging can be considered a two-stage approach [5], [6]. Human candidate extraction is the first stage, and the second stage involves classification. A multi-scale sliding window detector, thresholding segmentation, and background subtraction were used for candidate extraction [7]. The multi-scale sliding window detector is one of the most popular techniques for pedestrian detection in both visual camera and thermal camera image processing [8], [9].

By applying thresholding segmentation or background subtraction, the computation time can be decreased through a reduction in several regions [10], [11]. In thermal imaging applications, the thresholding segmentation relies on the premise that the human (or object) emits more heat and is warmer than the background. The distinction between objects and backgrounds is typically substantial, making this method ineffective in detecting humans who are cooler than the background, such as during the summer or daytime [12]. However, it demands considerable computational power and memory resources and usually requires a fixed camera and a static background [13], [14], [15].

Many methods based on deep learning have been applied to both visual and thermal images to enhance the accuracy of detection systems [16], [17]. With the recent advances of convolutional neural network (CNNs) in the computer vision community, they are attracting the attention of researchers exploring how to effectively extract information from both visual and thermal images [18], [19]. Deep learning models based on CNN architectures were applied to replace traditional classifiers in the classification stage. These approaches yielded significant results.

Trinh and Kim [9] used a multi-scale sliding window approach with image pyramids, combined with CNN, for binary classification in real-time pedestrian detection using thermal imaging. An effective human detection system in thermal imaging, based on a combination of background modeling and CNN, was proposed by Shahid *et al.* [20]. The authors presented a method for detecting pedestrians in thermal images where adaptive fuzzy C-means clustering and CNNs were utilized [21]. The adaptive fuzzy C-means was used as a thresholding segmentation approach to determine candidate regions. The CNN model was then applied as a binary classification. In study [22], a combination of two techniques, namely K-means clustering and the tiny you only look once, version 3 (YOLO v3), is employed for processing thermal images. This approach follows a two-step procedure. Initially, anchor boxes are created using the K-means technique. These anchor boxes play an important role in identifying the boundaries of objects. Subsequently, the tiny YOLO v3 model is employed to effectively forecast the boundary boxes of the identified objects, using the anchor boxes as references.

Recently, end-to-end deep learning has been widely and successfully applied to human detection in thermal imaging. Ivašić-Kos *et al.* [23] proposed using the YOLO model to locate humans in thermal images. Wang and Hosseinyalamdary [24] proposed using the RetinaNet method to address the issue of detecting humans in thermal images. They use additional information from temporal components in the video as compared to still images, resulting in improved human detection outcomes. In [25], a pixel-wise method based on CNN is proposed to address the challenge of detecting humans. The results presented in this paper are compared with those of the five traditional and most effective approaches. Li [26] focused on the detection of pedestrians within thermal images by employing the YOLO v3 model. The authors introduced a strategy that utilizes the strengths of YOLO v3, a deep learning architecture, to precisely identify pedestrians in infrared images. The study [27] introduced an innovative strategy for detecting humans through the utilization of YOLO v5 and transfer learning. The research centers on the utilization of thermal image data obtained from unmanned aerial vehicles (UAVs) to enhance surveillance capabilities. Akshatha *et al.* [28] investigated Faster R-CNN and single shot multibox detector (SSD) to detect humans in aerial thermal images.

In this study, we propose a comparison of different feature learning approaches based on deep learning for detecting humans in thermal images. Specifically, we aim to enhance the performance of the human detection system by considering the effectiveness of the Mask R-CNN model. The Mask R-CNN is an extension of the Faster R-CNN, which not only extracts bounding boxes containing human objects but also identifies segmentation masks around the subjects. We also explore and compare the effectiveness of various CNNs for feature learning, such as ResNet50 V1, ResNet101 V1, and Inception ResNet V2, within the Mask R-CNN framework. We evaluated the experiment using our human dataset for thermal imaging. The results show that the proposed method obtains high performance in detecting people in thermal imaging.

2. PROPOSED METHOD

In this research, thermal images are analyzed using the Mask R-CNN model [29] with varying CNN backbone structures to identify human subjects. The flowchart representation of the new approach is illustrated in Figure 1. The Faster R-CNN model architecture that has been further improved in terms of both training and detection speed is suggested in [30].

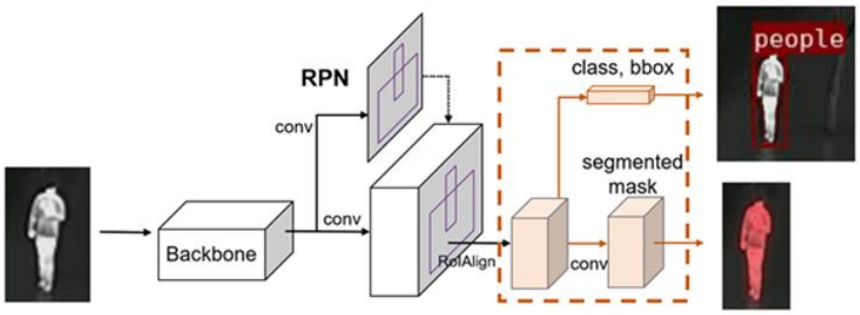

Figure 1. Overview of the Mask R-CNN model for human detection in thermal images [29]

This Faster R-CNN consists of two modules:

- Regional proposal network (RPN): A CNN network that suggests regions and object types to consider within the region.
- Fast R-CNN: A CNN network that extracts features from region proposals and produces bounding boxes and labels.

Both modules operate on the same output of a deep CNN. The RPN works by taking the output of a pretrained deep CNN and feeding the feature map into a small network that generates multiple region proposals and prediction labels. Region proposals are bounding boxes based on predefined anchor boxes or shapes designed to expedite and enhance region recommendations. The prediction for the label is expressed in binary, indicating whether an object is present or not in the region proposal. The first stage employs a deep CNN to generate a feature map. Unlike Fast R-CNN, this architecture does not directly generate RoIs on the feature maps; instead, it employs the feature maps as input to the RPN network to detect region proposals. Simultaneously, the feature maps are fed into the classifier to classify objects within the region proposals identified by the RPN network.

Building upon Faster R-CNN, Mask R-CNN introduces a third parallel branch for predicting object masks. Mask detection involves a fully connected network applied to each RoI. As pixel-level segmentation demands finer alignment compared to bounding boxes, Mask R-CNN enhances the RoI Pooling layer to achieve more precise mapping of RoIs to the regions in the original image. Mask R-CNN applies the same two-stage process, with the first stage identical to Faster R-CNN (that's RPN). In the second stage, in parallel with layer prediction and box offset, Mask R-CNN also outputs a binary mask for each RoI. The Mask R-CNN segmentation will use the result of the "heat map" and from here, apply deconvolution and unpooling to obtain the mask on the original image. Through deconvolution and unpooling, we can build a predictive partition on the original image for all classes of objects. This is also the output for the object partition block.

In addition to Faster R-CNN, the Mask R-CNN incorporates a significant improvement by replacing the RoI Pooling block with a module called RoI Align. This alteration plays a pivotal role in enhancing the accuracy of Mask R-CNN. Both RoI Pooling and RoI Align share the primary objective of standardizing the size of the region of interest for the subsequent layers. RoI Pooling [30] involves generating a compact feature map (e.g., 7×7) from each RoI. On the other hand, RoI Align is devised to address the localization errors inherent in RoI Pooling. RoI Align mitigates severe quantization effects; for instance, it employs x/16 instead of [x/16] to avoid coarse quantization. This precise alignment ensures that the extracted features align accurately with the input pixels.

2.1. Backbone CNNs

We used the network architecture for Mask R-CNN based on how to select the network for the feature extraction (backbone) corresponding to the top of the network corresponding to Figure 2. The backbones we use in Mask R-CNN are based on well-known CNN models. These CNN models serve as feature extractors from the input image. CNN has proven effective and significant feature learning.

We investigate and compare the effectiveness of some state-of-the-art deep CNN models as the backbone for human detection in thermal images, such as ResNet50 [31], ResNet101 [31], and Inception ResNet [32]. These architectures have been determined to have high results in object recognition and are widely used by other studies.

We implement the Mask R-CNN model based on the TensorFlow object detection framework on Google Colab. The process involves several steps, including setting up the environment, preparing the dataset, configuring the model, and training the model.

- Step 1 (environment setup): Install the TensorFlow object detection API
- Step 2 (dataset preparation): For training Mask R-CNN model based on TensorFlow object detection API, we initially label humans in our infrared images using the Labelme tool, resulting in *.json* files containing labeled humans. Our dataset has been annotated with bounding boxes and masks. The next step involves converting the *.json* annotation files to COCO format, which includes images with humans labeled as "people", along with bounding boxes and masks. Because the object detection API uses the TFRecord file format, we proceed to convert the COCO files into TFRecord format.
- Step 3 (model configuration): We download and install pre-trained backbone models of Mask R-CNN based on the COCO 2017 dataset and their configuration files. The backbone models such as ResNet50-V1, ResNet101-V1, and Inception ResNet-V2 are used in our experiments. We adjust the configuration file with *num_classes* is set to 1 (only detect human label in thermal image) to match our requirements. The other parameters and configuration settings are used by default in the configuration file.
- Step 4: Train our object detection model based on TensorFlow.

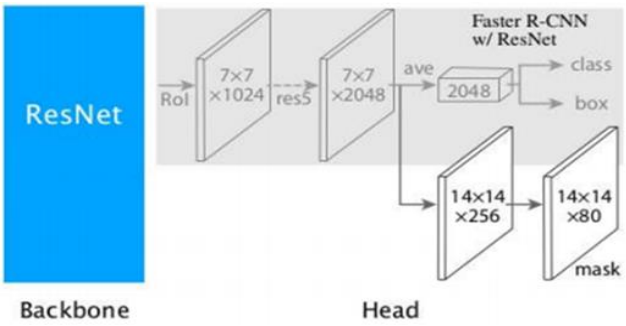

Figure 2. Mask R-CNN backbone architecture

3. RESULTS AND DISCUSSION

3.1. Dataset and experiment setup

We evaluate the performance of our detection system using a thermal dataset collected by the research team of Chonnam National University [7], [9]. The data is acquired using an Argo S thermal imaging camera operating at a resolution of 640×480, capturing frames at a rate of 31 frames per second. This camera is placed at a height of about 10~12 m from the ground (on the third floor of the building), as shown in Figure 3. The database was captured at various points around the Gwangju campus of Chonnam National University (CNU). It contains recordings from different locations and features varying sizes of moving objects, with the majority being quite small. The videos were taken outdoors during the summer months (July to September), when the temperatures typically range from 18 °C to 30 °C on average.

When employing thermal cameras, objects with higher temperatures than the background will appear brighter, while objects with lower temperatures will appear darker. During the night, as the body temperature is typically higher than the background, humans are seen as white due to this phenomenon known as polarity inversion. To address this issue, our newly devised method is designed to effectively manage such scenarios. Figure 4 shows examples of thermal images in our dataset.

In the experiments, we evaluate the outcomes of the proposed model by randomly selecting 540 thermal images from 3 videos. The training dataset comprises 448 images, while the test dataset comprises 92 images. In our dataset, several challenges impede accurate detection, as shown in Figure 4 such as occlusion (where humans can be partially obscured by tree branches, or when a car passes by and blocks the view, or is obstructed by other obstacles), various-sized objects (the humans in the image have varying sizes, especially most of them are small-sized and far from the camera's position), reflections (arising from varying surface

emissivity can distort thermal patterns, leading to inaccuracies), and so on. There is a significant amount of diversity in situations involving shifts in weather and lighting conditions, fluctuations in body and background temperatures, and sudden changes in intensity. This event takes place when the car enters the frame. At that moment, the camera automatically standardizes the thermal measurements for all areas within the scene and converts them into grayscale. Consequently, the vehicles appear brighter and cause the surrounding environment to become darker than its typical state. This results in a sudden shift in the frame's intensity, making it darker compared to the preceding frame. So, our dataset is a very challenging dataset.

Figure 3. Example of the thermal camera and location to record thermal videos [7], [9]

Figure 4. Examples of our thermal dataset

In our experiments, the size of each input image used to train the model was 640×480 pixels. Subsequently, the input images are passed through the RoI Pooling layer to be resized to 1024×1024 pixels. The input images are all maintained at a fixed size so that they can be grouped into multiple batches, thereby speeding up the training process.

To identify areas in the image that are likely to contain humans (regions of interest or candidates), we employ the intersection over union (IoU) metric for evaluation. In this study, we consider a region as a valid RoI (candidate) only when its IoU is greater than or equal to 0.5. If not, we exclude that region. This process is applied to all regions, resulting in a selection of regions where the IoU is greater than 0.5. The model's effectiveness is evaluated by calculating recall, precision, and F1-score based on these frames.

3.2. Results analysis

In this section, we evaluate the effectiveness of the Mask R-CNN algorithm for detecting humans in thermal imaging, utilizing various CNN-based backbones. We focus on the challenging conditions of thermal imaging for human detection methods, such as detecting small-sized objects, background similarity, and occlusion. Table 1 presents the performance comparison of the results of Mask R-CNN with various backbones. Based on Table 1, we observe that the Mask R-CNN model achieves a precision rate of over 96%, indicating a reduced likelihood of misrecognizing objects, which corresponds to lower noise detection. However, the recall rate is less satisfactory, attributed to challenges such as very small size of human

subjects, occlusion by tree leaves, and the similar intensity between humans and the background in the images, leading to a higher likelihood of the model overlooking them. Notably, ResNet50-V1 stands out with the highest recall at 79.33%, indicating its effectiveness in detecting humans. ResNet101-V1 demonstrates exceptional precision at 98.9%, implying accurate human predictions. Meanwhile, Inception ResNet-V2 obtains lower recall, precision, and F1-score. When considering the F1-score, which balances precision and recall, ResNet50-V1 maintains a solid balance with an F-value of 87.85%. These results emphasize the trade-offs and strengths of each model, offering valuable insights into their suitability for object detection tasks. Finally, our investigation revealed that the Mask R-CNN model with the ResNet50-V1 backbone outperformed others in accuracy, achieving a precision rate of 98.41% and a recall of 79.33%, with an F1- score of 87.85%. Figure 5 illustrates the comparison of the three backbone models for thermal imaging based on Mask R-CNN. Figure 5(a) represents the input thermal image. Figure 5(b) shows the results of Mask R-CNN using ResNet 50-V1. Figure 5(c) shows the results of Mask R-CNN using ResNet101-V1. Figure 5(d) shows the results of Mask R-CNN using Inception ResNet-V2.

Table 1. Performance results and computation cost of backbones in the Mask R-CNN model.

Backbones	Recall (%)	Precision (%)	F1– score (%)	Model training time (s)	Time to test each frame (s)
ResNet50-V1	79.33	98.41	87.85	1615.15	0.27
ResNet101-V1	75.16	98.90	85.41	2442.38	0.27
Inception ResNet-V2	72.59	96.38	82.81	3443.25	0.68

In addition, we compared training and testing times for each backbone model used. Table 1 also presents a comparison of the processing times of the backbones in the Mask R-CNN. We observed that ResNet50-V1 emerges as the most efficient in terms of training time, requiring only 1615.15 seconds to complete. ResNet101-V1 and Inception ResNet-V2 follow, with training times of 2442.38 seconds and 3443.25 seconds, respectively. When it comes to analyzing frames, both ResNet50-V1 and ResNet101-V1 exhibit the same processing speed, taking 0.27 seconds per frame. In contrast, Inception ResNet-V2 is relatively slower, needing 0.68 seconds for each frame. These results offer valuable insights for choosing a model based on computational efficiency, with ResNet50-V1 showcasing superior training speed and real-time frame analysis.

We also compare the performance of Mask R-CNN models with the SSD approach [28] on our dataset, as shown in Table 2. We realize that the Mask R-CNN variants significantly outperform the SSD using backbone ResNet50-V1 methods across all metrics. The Mask R-CNN with ResNet50-V1 demonstrates the best performance, with the highest F1-score and excellent precision, indicating a high rate of accurate detections and a low rate of false positives. Although the Mask R-CNN with ResNet101-V1 and Inception ResNet-V2 have slightly lower recall and F-values, they still maintain high precision and considerably outperform the SSD approach.

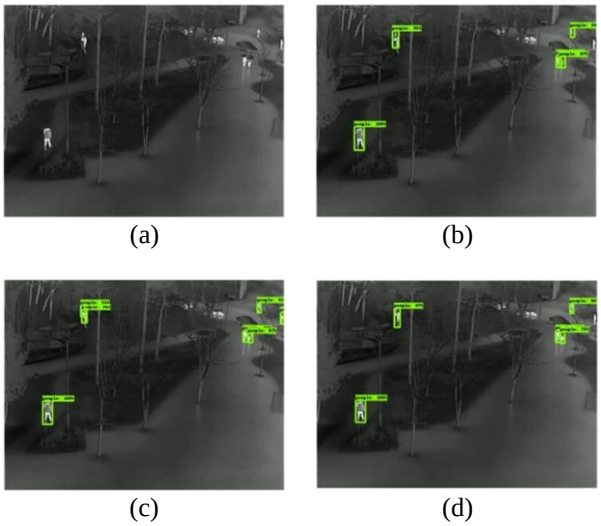

Figure 5. Comparison results with different CNN backbones: (a) input image, (b) result based on the ResNet 50-V1, (c) result based on the ResNet101-V1, and (d) result based on the Inception ResNet-V2

Table 2. Performance comparisons of various methods in our dataset

Approaches	Recall (%)	Precision (%)	F1– score (%)
Mask R-CNN with ResNet50-V1	**79.33**	**98.41**	**87.85**
Mask R-CNN with ResNet101-V1	75.16	98.90	85.41
Mask R-CNN with Inception ResNet-V2	72.59	96.38	82.81
SSD with ResNet50-V1	66.94	79.52	72.68

While the study presents a comprehensive analysis of Mask R-CNN backbones, it acknowledges limitations like the lower recall rate in complex imaging scenarios. In some cases, the objects have the same intensity as the background, leading to false negatives (missed detections). When people are close to each other or occluded by trees or other objects, parts of these objects are lost. This can make it challenging for the algorithm to detect and separate humans, potentially reducing the recall rate. The further away a person is from the camera, the smaller the object size becomes, and the less detail there is in the image for the algorithm to analyze, making accurate detection more difficult. Improving recall often comes at the cost of reducing precision, so there is usually a trade-off to be managed in these systems. The intricacies of these scenarios necessitate further, more detailed research to enhance detection accuracy. These limitations are illustrated in Figure 6 and will be considered in future work.

Figure 6. Examples of miss detection errors of our method in shown in the red ellipse regions

4. CONCLUSION

This study employs the Mask R-CNN framework, coupled with a comparison of backbones including ResNet 50-V1, ResNet 101-V1, and Inception ResNet-V2 to effectively detect humans in thermal imagery. We investigate human detection using a thermal database recorded at the Gwangju campus of Chonnam National University. This method shows great potential for enhancing campus security, providing an important capability for maintaining a secure environment through continuous monitoring. Furthermore, our approach can be extended to other application domains, such as search and rescue operations, where it has the potential to detect individuals in challenging environments. It can also be applied to critical areas like border control. Our model has demonstrated favorable accuracy results, as described in the experimental results section. However, there are certain limitations within this topic, outlined as follows: i) In frames where numerous objects are present, or objects are obscured by trees or exhibit lower temperatures than the background, the model does not achieve optimal results; ii) Dataset limitations: the actual dataset remains constrained; therefore, there is a necessity to amass more data and provide accurate labeling to enhance the model's training effectiveness. The dataset has remained relatively unchanged, consisting solely of three videos.

Our next goal is to expand the database by collecting data from various other sources that exhibit high variability. This approach will enable us to train the model more efficiently. Additionally, we plan to employ methods for enhancing video processing. Rather than treating each frame individually, we intend to explore the interframe relationships within videos.

ACKNOWLEDGEMENTS

This work was partly supported by Saigon University. The authors acknowledge Professor Jin Young Kim of the Intelligent Electronics Lab at Chonnam National University for providing the thermal dataset for research purposes and for his support.

REFERENCES

[1] W. Chen, Y. Zhu, Z. Tian, F. Zhang, and M. Yao, "Occlusion and multi-scale pedestrian detection A review," *Array*, vol. 19, Art. no. 100318, Sep. 2023, doi: 10.1016/j.array.2023.100318.

[2] M. Kristo, M. Ivasic-Kos, and M. Pobar, "Thermal object detection in difficult weather conditions using YOLO," *IEEE Access*, vol. 8, pp. 125459–125476, 2020, doi: 10.1109/ACCESS.2020.3007481.

[3] N. Bustos, M. Mashhadi, S. K. Lai-Yuen, S. Sarkar, and T. K. Das, "A systematic literature review on object detection using near infrared and thermal images," *Neurocomputing*, vol. 560, Art. no. 126804, Dec. 2023, doi: 10.1016/j.neucom.2023.126804.

[4] H. Tan, D. Ou, L. Zhang, G. Shen, X. Li, and Y. Ji, "Infrared sensation-based salient targets enhancement methods in low-visibility scenes," *Sensors*, vol. 22, no. 15, p. 5835, Aug. 2022, doi: 10.3390/s22155835.

[5] J. Cao, Y. Pang, J. Xie, F. S. Khan, and L. Shao, "From handcrafted to deep features for pedestrian detection: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 9, pp. 4913–4934, Sep. 2022, doi: 10.1109/tpami.2021.3076733.

[6] K. Piniarski, P. Pawłowski, and A. Dąbrowski, "Tuning of classifiers to speed-up detection of pedestrians in infrared images," *Sensors*, vol. 20, no. 16, p. 4363, Aug. 2020, doi: 10.3390/s20164363.

[7] T. D. Trinh, X. Ma, and J. Y. Kim, "Improved running gaussian average for background subtraction in thermal imagery," *Journal of Korean Institute of Information Technology*, vol. 15, no. 7, pp. 101–117, Jul. 2017, doi: 10.14801/jkiit.2017.15.7.101.

[8] J. Baek, S. Hong, J. Kim, and E. Kim, "Efficient pedestrian detection at nighttime using a thermal camera," *Sensors*, vol. 17, no. 8, Art. no. 1850, Aug. 2017, doi: 10.3390/s17081850.

[9] T. D. Trinh and J. Y. Kim, "Multi-scale pedestrian detection in thermal imaging using deep convolutional neural network and adaptive NMS," *The Journal of Korean Institute of Information Technology*, vol. 16, no. 9, pp. 85–94, Sep. 2018, doi: 10.14801/jkiit.2018.16.9.85.

[10] A. Sobral and A. Vacavant, "A comprehensive review of background subtraction algorithms evaluated with synthetic and real videos," *Computer Vision and Image Understanding*, vol. 122, pp. 4–21, May 2014, doi: 10.1016/j.cviu.2013.12.005.

[11] R. Kalsotra and S. Arora, "Background subtraction for moving object detection: explorations of recent developments and challenges," *The Visual Computer*, vol. 38, no. 12, pp. 4151–4178, Aug. 2021, doi: 10.1007/s00371-021-02286-0.

[12] E. Jeon *et al.*, "Human detection based on the generation of a background image by using a far-infrared light camera," *Sensors*, vol. 15, no. 3, pp. 6763–6788, Mar. 2015, doi: 10.3390/s150306763.

[13] B. Garcia-Garcia, T. Bouwmans, and A. J. Rosales Silva, "Background subtraction in real applications: Challenges, current models and future directions," *Computer Science Review*, vol. 35, Art. no. 100204, Feb. 2020, doi: 10.1016/j.cosrev.2019.100204.

[14] X. Zhang, P. Ye, H. Leung, K. Gong, and G. Xiao, "Object fusion tracking based on visible and infrared images: A comprehensive review," *Information Fusion*, vol. 63, pp. 166–187, Nov. 2020, doi: 10.1016/j.inffus.2020.05.002.

[15] T. Bouwmans, S. Javed, M. Sultana, and S. K. Jung, "Deep neural network concepts for background subtraction: A systematic review and comparative evaluation," *Neural Networks*, vol. 117, pp. 8–66, Sep. 2019, doi: 10.1016/j.neunet.2019.04.024.

[16] Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015, doi: 10.1038/nature14539.

[17] L. Alzubaidi *et al.*, "Review of deep learning: concepts, CNN architectures, challenges, applications, future directions," *Journal of Big Data*, vol. 8, no. 1, Mar. 2021, doi: 10.1186/s40537-021-00444-8.

[18] J. Kim, H. Hong, and K. Park, "Convolutional neural network-based human detection in nighttime images using visible light camera sensors," *Sensors*, vol. 17, no. 5, p. 1065, May 2017, doi: 10.3390/s17051065.

[19] D. Guan, Y. Cao, J. Yang, Y. Cao, and M. Y. Yang, "Fusion of multispectral data through illumination-aware deep neural networks for pedestrian detection," *Information Fusion*, vol. 50, pp. 148–157, Oct. 2019, doi: 10.1016/j.inffus.2018.11.017.

[20] N. Shahid, G.-H. Yu, T. D. Trinh, D.-S. Sin, and J.-Y. Kim, "Real-time implementation of human detection in thermal imagery based on CNN," *The Journal of Korean Institute of Information Technology*, vol. 17, no. 1, pp. 107–121, Jan. 2019, doi: 10.14801/jkiit.2019.17.1.107.

[21] V. John, S. Mita, Z. Liu, and B. Qi, "Pedestrian detection in thermal images using adaptive fuzzy C-means clustering and convolutional neural networks," in *2015 14th IAPR International Conference on Machine Vision Applications (MVA)*, IEEE, May 2015. doi: 10.1109/mva.2015.7153177.

[22] P. Talluri and M. Dua, "Low-resolution human identification in thermal imagery," in *2020 5th International Conference on Communication and Electronics Systems (ICCES)*, IEEE, Jun. 2020. doi: 10.1109/icces48766.2020.9138039.

[23] M. Ivašić-Kos, M. Krišto, and M. Pobar, "Human detection in thermal imaging using YOLO," in *Proceedings of the 2019 5th International Conference on Computer and Technology Applications*, in ICCTA 2019. ACM, Apr. 2019. doi: 10.1145/3323933.3324076.

[24] X. Wang and S. Hosseinyalamdary, "Human detection based on a sequence of thermal images using deep learning," *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives*, vol. 42, no. 2/W13, pp. 127–132, Jun. 2019, doi: 10.5194/isprs-archives-XLII-2-W13-127-2019.

[25] J. Park, J. Chen, Y. K. Cho, D. Y. Kang, and B. J. Son, "CNN-based person detection using infrared images for night-time intrusion warning systems," *Sensors*, vol. 20, no. 1, p. 34, Dec. 2019, doi: 10.3390/s20010034.

[26] W. Li, "Infrared image pedestrian detection via YOLO-V3," in *IEEE Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, IEEE, Mar. 2021, pp. 1052–1055. doi: 10.1109/IAEAC50856.2021.9390896.

[27] A. J. Mantau, I. W. Widayat, J.-S. Leu, and M. Köppen, "A human-detection method based on YOLOv5 and transfer learning using thermal image data from UAV perspective for surveillance system," *Drones*, vol. 6, no. 10, Art. no. 290, Oct. 2022, doi: 10.3390/drones6100290.

[28] K. R. Akshatha, A. K. Karunakar, S. B. Shenoy, A. K. Pai, N. H. Nagaraj, and S. S. Rohatgi, "Human detection in aerial thermal images using faster R-CNN and SSD algorithms," *Electronics*, vol. 11, no. 7, Art. no. 1151, Apr. 2022, doi: 10.3390/electronics11071151.

[29] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask R-CNN," in *2017 IEEE International Conference on Computer Vision (ICCV)*, IEEE, Oct. 2017. doi: 10.1109/iccv.2017.322.

[30] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017, doi: 10.1109/TPAMI.2016.2577031.
[31] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, IEEE, Jun. 2016, pp. 770–778. doi: 10.1109/CVPR.2016.90.
[32] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-ResNet and the impact of residual connections on learning," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, no. 1, Feb. 2017, doi: 10.1609/aaai.v31i1.11231.

BIOGRAPHIES OF AUTHORS

Tan Dat Trinh received a B.Sc. degree in mathematics and computer sciences from University of Natural Science – National University of HCM City, Vietnam in 2010. He also received Master of Engineering (M.Eng.) and Ph.D. degree in electronics and computer engineering from Chonnam National University, Korea in 2013 and 2017, respectively. He is currently a lecturer at Computer Science Department, Sai Gon University, Vietnam since 2019. He is also an AI researcher at Computer Science Laboratory, Sai Gon University since 2019. His research areas of interest include speaker recognition, speech signal processing, computer vision and pattern recognition. He can be contacted at email: trinhtandat@sgu.edu.vn.

Pham Cung Le Thien Vu received his B.Sc. (mathematics and information technology) in 2015 and is studying for master's degree of computer science from Ho Chi Minh University of Science, Vietnam. He is currently an AI and machine learning engineer at Heligate JSC company. He is also an AI researcher at Computer Science Laboratory, Sai Gon University, Vietnam since 2020. His research includes machine learning, deep learning, speaker recognition, computer vision, and natural language processing. He can be contacted at email: phamcunglethienvu@gmail.com.

Pham The Bao received his B.Sc. degree in Algebra from University of Natural Science – National University of HCM City, Vietnam in 1995. He also received MSc degree in Mathematical Foundation of Computer Science and Ph.D. degree in computer science from University of Natural Science – National University of HCM City, Vietnam in 2000 and 2009, respectively. He was a lecturer and professor in Department of Computer Science, Faculty of Mathematics Computer Science, University of Natural Science, Vietnam from 1995 to 2018. He is currently dean and professor at Computer Science Department, Sai Gon University, Vietnam since 2019. He has published over 50 papers in international journals and conferences. His research includes image processing, pattern recognition and intelligent computing. He can be contacted at email: ptbao@sgu.edu.vn.