OBJECT DETECTION IN UAV-BORNE THERMAL IMAGES USING BOUNDARY-AWARE SALIENCY MAPS

Minglei Li[1, *], Xingke Zhao[1], Jiasong Li[1], Daiyin Zhu[1]

[1] College of Electronic and Information Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing, China – (minglei_li, zxk313, jeasonlee, zhudy)@nuaa.edu.com

Commission II, WG II/III

KEY WORDS: Thermal Image, Deep Learning, Object Detection, Saliency Map, YOLOv3

ABSTRACT:

In this paper, we propose a method of object detection based on thermal images acquired from unmanned aerial vehicles (UAV). Compared with visible images, thermal images have lower requirements for illumination conditions, but they have some problems, such as blurred edges and low contrast. To address these problems, we propose to use the saliency map of thermal images for image enhancement as the attention mechanism of the object detector. In the paper, the YOLOv3 network is trained as a detection benchmark and BASNet is used to generate saliency maps from the thermal images. We fuse the thermal images with their corresponding saliency maps through the pixel-level weighted fusion method. Experiment results tested on real data have shown that the proposed method could realize the task of object detection in UAV-borne thermal images. The statistical results show that the average precisions (AP) of pedestrians and vehicles are increased by 4.5% and 2.6% respectively, compared with the benchmark of the YOLOv3 model trained on only the thermal images. The proposed model provides reliable technical support for the application of thermal images with UAV platforms.

1. INTRODUCTION

Image-based object detection techniques have been widely used in several applications, such as environmental monitoring, emergency management, and traffic survey. However, most of the previous works focus on visible images, which might be influenced by illumination. On the contrary, thermal images have the capability of observing objects at night or in bad lighting conditions. Hence, more researchers are exploring the potential of using thermal images to build intelligent systems (Portmann et al., 2014; Wang, Bai, 2019; Li et al., 2019; Sun et al., 2019). Compared with visible images, thermal images have some defects, such as low contrast, edge blur, and strong noise, which make them less distinguishable. In addition, UAV jitter can cause image blur. To address the challenge in UAV-borne thermal images, we propose the use of boundary-aware saliency maps to enhance the data.

The purpose of salient object detection is to highlight the most obvious objects in an image. These methods can guide machine vision systems to allocate limited computing resources to a few salient regions (Klein, Frintrop, 2011; Zhang et al., 2015; Cheng et al., 2015; Qin et al., 2019). Object-driven salient detection algorithms mainly focus on the image content based on the task requirements, and the results of saliency detection are determined by corresponding tasks. In recent years, deep convolutional neural networks (CNN) (Yann et al., 1998) have been used for salient object detection and achieved state-of-the-art performance. In this paper, we adopt the boundary-aware salient object detection network BASNet (Qin et al., 2019) to generate saliency maps from thermal images. Then, the saliency map is fused with the original thermal image. We suppose that the fusion images maintain texture information and clear boundaries of objects, which can enhance the recognition capability of the trained models using the fusion images.

Our reference model is established by training a state-of-the-art object detector YOLOv3 (Redmon, Farhadi, 2018). As the original thermal image is a single channel grayscale image, we first convert it to an indexed visible image (with 3 channels), simulating the hues of glowing iron, i.e. iron red color model. Then, we can adopt the existing deep CNNs to generate the saliency map. Our training data of saliency maps of thermal images are manually prepared by the *Labelme* toolbox (Kentaro, 2016), which provides tools to generate pixel level image annotation.

The main contributions of this paper are as follows:

(1) This paper presents a method of fusing the thermal images with their corresponding saliency maps. Based on the fusion images, the trained deep learning model demonstrates the effect of saliency maps on improving object detection performance of thermal images from the perspective of UAV.

(2) We release a dataset of thermal images with the annotation information, which is useful for the research of deep learning techniques based on thermal images. The data can be found in: https://drive.google.com/drive/folders/1vCxXsKnK3dVB-bkT6XLbbQF7YTdS2CR0?usp=sharing. We provide saliency detection benchmarks on it using state-of-the-art networks.

2. RELATED WORK

To the best of our knowledge, there are few papers discussing deep learning methods with UAV-borne thermal images to detect objects. In the following, we review related literatures in the aspects of object detection and saliency maps.

2.1 Object Detection

Over the past 20 years, a great deal of research has been devoted to pedestrian and vehicle detection from visible images. Focusing on the problem of self-occlusion in the field of human motion

[*] Corresponding author

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLIII-B2-2020-1233-2020 | © Authors 2020. CC BY 4.0 License.

tracking, Yu et al. (2010) deal with an algorithm for detecting the pedestrian limbs self-occlusion probability model. This algorithm uses the Markov model and the ellipse skin color model to change the detection of pedestrian limbs self-occlusion to the calculation of the self-occlusion state transition probability. The result of the experiment shows that the algorithm has higher accuracy. Dollar et al. (2009) put forward the method of combining Integral Channel Feature with Boosting algorithm to improve the effect of vehicle detection. Compared with traditional detection algorithms, CNN has made a significant breakthrough in object detection in recent years. Object detection algorithms based on CNN are mainly divided into two categories: One-stage and Two-stage methods of object detection. The main difference between them is whether there is a cascade module that extracts region proposals. The representatives of two-stage object detection algorithms are the R-CNN series of detection algorithms (Girshick et al., 2014; Girshick, 2015; Ren et al., 2017). They can make the network to detect objects in the suspected object area by using the cascade module. The cascaded module will increase the complexity of the model while increasing accuracy. It is lower than the one-stage detection algorithm in the detection speed and is not suitable for real-time object detection on UAV. Although the one-stage object detection algorithm performs poorly in detection accuracy, its detection speed is very fast. The most representative of them is the YOLO series of object detection algorithms (Redmon, Farhadi, 2018; Redmon et al., 2016; Redmon et al., 2017). The YOLOv3 algorithm divides images into S×S grids, and each grid is responsible for object detection whose center is in this grid. Detection and recognition are completed at the same time using regression methods. Based on this, we use the YOLOv3 algorithm as the basic model for pedestrian and vehicle object detection and recognition in UAV thermal images.

In recent years, more and more researches focus on the use of thermal images to realize the effective detection of pedestrians and vehicles. Zhang et al. (2010) proposed an automatic visual-thermal image sequence registration method based on co-motion and the results showed that the proposed algorithm carried out precise image registration under the change of image rotation, translation, scaling and viewing angle. Li et al. (2019) introduced a neural network for light perception, which adaptively fused the optical and thermal subnetworks, and adopted a weighted scheme to fuse the results according to the light conditions. A regional reconstruction network was introduced by Xu et al. (2017), and CNN was used to model the relationship between visible and thermal data, and then these features were input into the multi-scale detection network for robust object detection. In our context, however, we use only thermal images to design a general detection framework that works for day and night.

2.2 Saliency Detection

The purpose of salient object detection is to highlight the most obvious objects in an image. It can guide machine vision systems to allocate limited computing resources to a few salient regions, which provides great convenience for subsequent visual processing. Related research can be divided into two categories: data-driven and object-driven saliency detection. Data-driven salient region detection algorithms mainly focus on the visual stimulus caused by the underlying features of the image. These algorithms are driven by internal data and independent of the object task. On the contrary, object-driven salient detection algorithms mainly focus on the image content based on the task requirements, and the results of saliency detection are determined by corresponding tasks.

Laurent et al. (1998) first proposed the visual attention model. In this model, multiscale image features are combined into a single topographical saliency map. A dynamical neural network then selects attended locations in order of decreasing saliency. Hou et al. (2007) proposed a spectral residual approach. By analyzing the log-spectrum of an input image, they extract the spectral residual of an image in the spectral domain and design a fast method to construct the corresponding saliency map in the spatial domain. Using deep learning techniques to generate a visual saliency map becomes a trend in recent studies. A novel super pixel-wise convolutional neural network approach, called SuperCNN, is proposed by He et al. (2015) to learn the internal representations of saliency in an efficient manner. In contrast to the classical convolutional networks, SuperCNN is able to learn the hierarchical contrast features, and saliency can be detected independent of region size by utilizing a multi-scale network structure. Hou et al. (2017) propose a new method for saliency detection by introducing short connections to the skip-layer structures within the Holistically-Nested Edge Detector architecture. This framework provides rich multi-scale feature maps at each layer, a property that is critically needed to perform segment detection. In this paper, we used the most advanced network BASNet to generate saliency maps from thermal images and to build a benchmark of saliency maps.

3. ALGORITHM

3.1 Baseline for Pedestrian and Vehicle Detection in Thermal Images using YOLOv3

We adapt the YOLOv3 network for the task of objection detection. Specifically, our current targets are pedestrians and vehicles, and the network we used has been pre-trained on the COCO dataset (Lin et al., 2014). YOLOv3 adopts DarkNet53 with higher accuracy as the image feature extraction network and designs a multi-scale detection structure, which has good adaptability to small objects suitable for UAV-borne data. As shown in Figure 1, we fine-tuned the traditional YOLOv3 network on the original thermal images to generate the retrained model that serves as the benchmark, then the retrained model on only saliency maps and the model on only fusion images are compared with the benchmark.

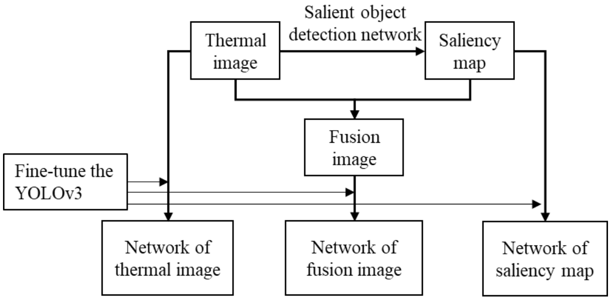


Figure 1. Comparison mechanism between the traditional YOLOv3 model, saliency map detection model, and the proposed model.

3.2 Saliency Map Generation

Deep CNN has been used in salient object detection and achieved good performance. But most previous works are focused on the accuracy of areas, not the quality of boundaries. As we propose to enhance the thermal image from the perspective of UAV by using the saliency map, the boundary of the salient object will have a great impact on the image enhancement effect. Therefore, we adapt the BASNet network, which is more concerned with boundary quality, as the base network for generating saliency maps.

As shown in Figure 2, the architecture of the salient object detection network BASNet is composed of (1) **a prediction module** and (2) **a residual refinement module**.

The prediction module is similar to U-shape-Net (Olaf et al., 2015). First, it has an encoder phase consisting of a convolution layer and six basic res-blocks adopted from ResNet-34. As symmetry, the module then runs a decoding phase, which also has six stages. Each stage consists of convolution layers followed by a batch normalization (BN) and a ReLU activation function. The input of each stage is the concatenated feature maps of the up-sampled output from its previous stage and its corresponding stage in the encoder. This module yields a coarse saliency map, where the boundaries of objects are inaccuracy.

Then, **the residual refinement module** refines the saliency map of the prediction module by learning the residuals between the predicted saliency map and the ground truth. This model also has an encoder phase and a decoder phase. Different from the predict module, both encoder and decoder have four stages. Each stage only has one convolution layer followed by a BN and a ReLU. Non-overlapping max pooling and bilinear interpolation are utilized in the res-sampling stages. The final output is the refined saliency map, which will be used to fuse with the original thermal image.

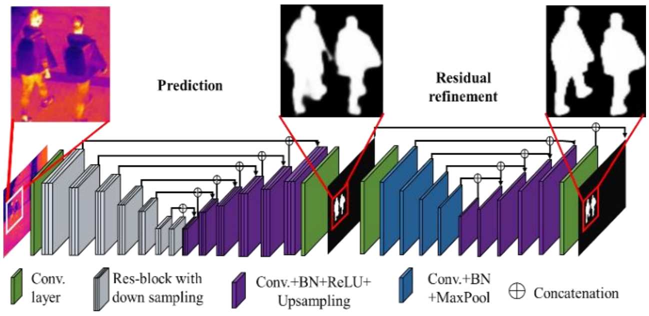


Figure 2. Architecture of the boundary-aware saliency map detection network BASNet.

Different from other prediction networks, BASNet uses the mixed loss of Binary Cross Entropy (BCE), Structural SIMilarity (SSIM) and Intersection-over-Union (IoU) to design the loss function for each layer, so the network pays more attention to boundary quality instead of only focusing on regional accuracy. The loss is defined as:

$$L = L_{bce} + L_{ssim} + L_{iou} \tag{1}$$

L_{bce} denotes BCE loss, corresponding to pixel level supervision:

$$L_{bce} = -\sum_{(r,c)}[G(r,c)\log(S(r,c)) + (1 - G(r,c))\log(1 - S(r,c))] \tag{2}$$

where $G(r,c) \in \{0,1\}$ is the ground truth label of the pixel (r,c), and $S(r,c)$ is the predicted probability of being salient objects. L_{ssim} denotes SSIM loss, corresponding to the supervision at the patch level:

$$\frac{(2\mu_x\mu_y+C_1)(2\sigma_{xy}+C_2)}{(\mu_x^2+\mu_y^2+C_1)(\sigma_x^2+\sigma_y^2+C_2)} \tag{3}$$

where $x = \{x_j: j = 1,\cdots,N^2\}$ and $y = \{y_j: j = 1,\cdots,N^2\}$ are the pixel values of two corresponding patches (size: $N \times N$) cropped from the predicted probability map S and the binary ground truth mask G respectively. μ_x, μ_y and σ_x, σ_y are the mean and standard deviations of x and y, σ_{xy} is their covariance, and $C_1 = 0.01^2, C_2 = 0.03^2$ are used to avoid dividing by zero.

L_{iou} denotes IoU loss, corresponding to the supervision at the level of map:

$$\frac{\sum_{r=1}^{H}\sum_{c=1}^{W} S(r,c)G(r,c)}{\sum_{r=1}^{H}\sum_{c=1}^{W}[S(r,c)+G(r,c)-S(r,c)G(r,c)]} \tag{4}$$

where $S(r,c)$ and $G(r,c)$ are consistent with those represented in L_{ssim}.

3.3 Fusion of Thermal Images with Saliency Maps

The saliency map serves as an attention mechanism, but it discards all textural information. We augment the thermal images with the corresponding saliency maps to create a new fusion image. The 3-channels of the fusion image are built by combining the saliency map values and the values in thermal image channels. As shown in Figure 3, the fusion image strengthens the salient parts of the image, while retaining the textural information.

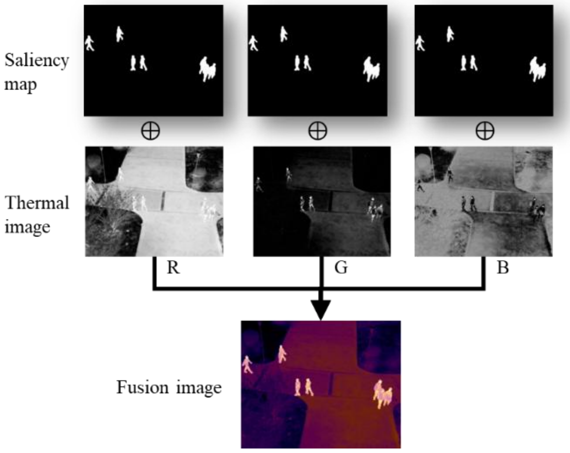


Figure 3. The fusion of the thermal image with the corresponding saliency map.

4. EXPERIMENTS

4.1 Datasets and Evaluation Protocols

In order to train a deep neural network, a large number of data samples are needed. However, at present, there is no publicly available thermal dataset for pedestrians and vehicles from the perspective of UAV. In addition, salient object detection also requires pixel-level annotations of the salient objects. Therefore, we made the pedestrian and vehicle dataset based on thermal images under UAV, and we made it publicly available to facilitate further research on multispectral pedestrian and vehicle saliency detection technology. It is worth noting that the original image received by the thermal camera has only brightness and is a single-channel grayscale image. In order to facilitate the research of pedestrian and vehicle object detection, thermal images received are converted into a three-channel pseudo-color image of RGB format after temperature mapping. The pixel with value 0 in single-channel grayscale images is mapped to blue and the pixel with value 255 in single-channel grayscale images is mapped to red, with a smooth gradation in the middle, that uses the warm and cold tones of the color to display low and high temperature areas.

The imaging system is built on a DJI M600Pro UAV carrying a FLIR vue pro thermal camera. It captured data in the daytime and at night. The dataset contains 2975 thermal infrared images, including 4768 pedestrian instances and 3856 vehicle instances. Labelme toolbox is used to manually annotate these images to generate the required training data and evaluation data. The

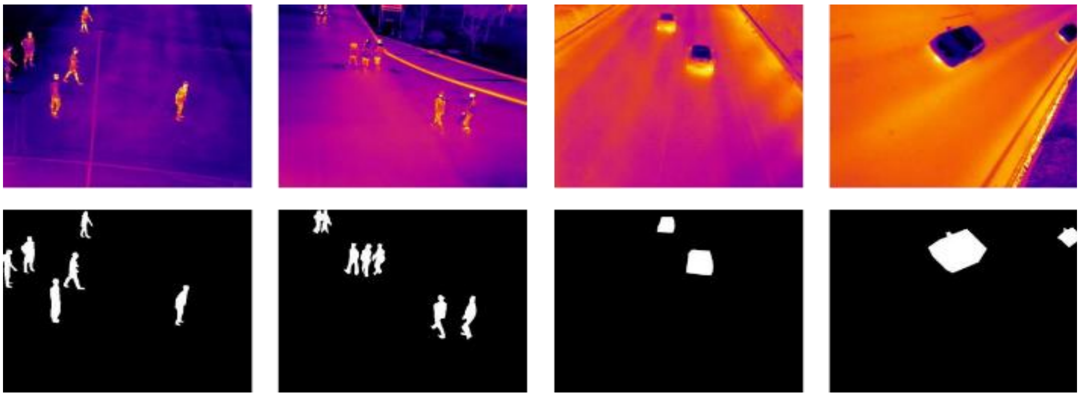
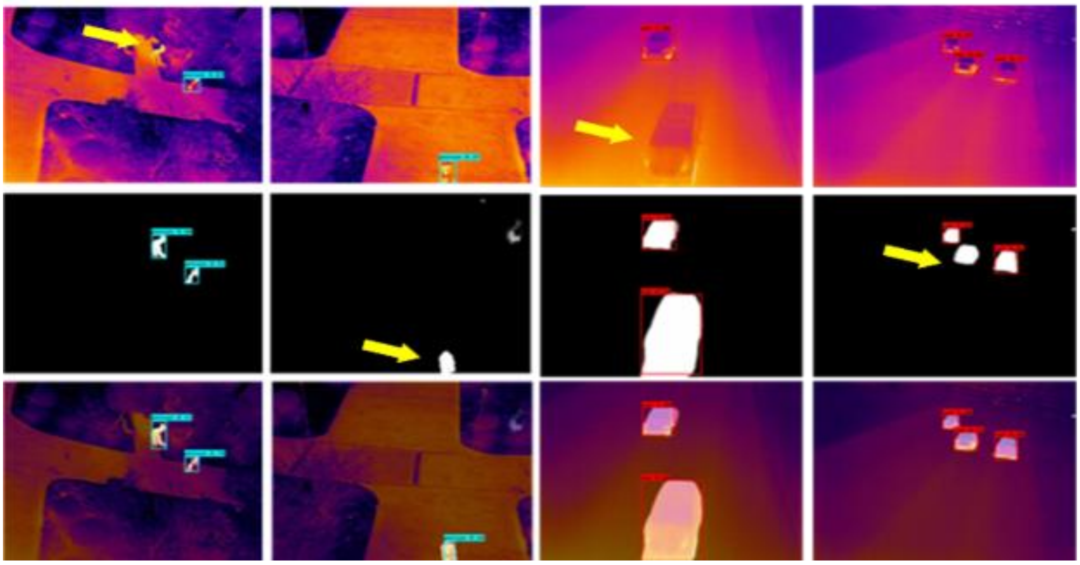


Figure 4. Sample annotations from pedestrian and vehicle thermal dataset. Top: Original images. Bottom: Pixel level annotations.



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

Figure 5. Example results of the detection results of different model trained only on (1st row) origianl thermal images, (2nd row) saliency maps, and (3rd row) fusion images. The arrows indicate some missing objects.

annotation here includes not only the border information of target locations and categories but also the pixel level annotation for saliency object detection. Even though the volume of the thermal image dataset is not huge, our networks can converge efficiently, as both YOLOv3 and the boundary-aware saliency detection network have pre-trained models for pedestrians and vehicles. Figure 4 shows some example images and annotations of the dataset we made.

In order to evaluate the results of pedestrian and vehicle detection, AP and frame per second (FPS) are used as the evaluation indexes of accuracy and speed respectively. In addition, F-measure (F_β) and Mean absolute error (MAE) are used to evaluate the saliency detection results of the model. Where, F_β is the weighted harmonic mean of precision and recall under the condition of non-negative weighted degree β, and the higher F_β is, the better the model is. The specific formula is as follows:

$$F_\beta = \frac{(1+\beta^2)Precision \times Recall}{\beta^2 Precision + Recall} \tag{5}$$

β^2 is generally 0.3.

MAE is to directly calculate the pixel error between the saliency map output by the model and its corresponding ground truth value:

$$MAE = \frac{1}{W \times H} \sum_{x=1}^{W} \sum_{y=1}^{H} |\bar{S}(x,y) - \bar{G}(x,y)| \tag{6}$$

Where, W and H are the width and height of the image, $\bar{S}(x,y)$ and $\bar{G}(x,y)$ are the pixel values of the output saliency map and its corresponding ground truth.

4.2 Implementation Details of Network Models

We train the YOLOv3 model on NVIDIA 1080ti GPU with 12GB video memory. The image size is adjusted from 640×512 to 416×416 by bilinear interpolation and then input into network models. We fine-tune the network for 100 epochs with a learning rate of 0.001 and batch size of 8, using original images, saliency images, and fusion images respectively. Besides, we augment the training images with random mirror flipping and random crops. The IoU threshold is set to 0.5, and the final prediction result is output after the non-maximum suppression (NMS) operation.

We use pixel-level labeled thermal images to train BASNet and maintains the same network architecture as in the original paper. In the training phase, the size of each image in the training set is first adjusted to 256 × 256, and the training set images are enhanced by random flipping and cropping. The weights of the ResNet-34 network are used to initialize the parameters of the feature extraction network, the decoding network is trained from 0, and the learning rate is 0.01. Without using a validation set and a batch size of 8, the loss function converges after 60,000 iterations, and the entire training process takes about 7 hours. In the test phase, input image size is also adjusted to 256 ×256, and then the images are input to the network to obtain predicted saliency maps. In the end, the down-sampled saliency map is resized to the size of the original input image. Both adjustments use bilinear interpolation.

5. RESULTS AND ANALYSIS

5.1 Detection Effect of Deep Saliency Network BASNet on Thermal Image Dataset

In order to provide effective support for the following saliency map study, the performance of BASNet in the annotated UAV thermal pedestrian and vehicle saliency dataset is evaluated. The evaluation results show that F_β is 0.767 and MAE is 0.008. It can be seen that the detection result of the model is very good.

5.2 Object Detection Analysis of Fusion Saliency Maps

In Figure 5, we have shown the detection results of the different models trained only on origianl thermal images (1st row), saliency maps (2nd row), and fusion images (3rd row). We can observe that the saliency maps indeed contribute to improved performance, as some missing objects in the thermal images and the saliency maps are found out in the fusion images. It can be seen in images (a) and (c) that a pedestrian and a track have been lost in the detection results in thermal images. The saliency maps (b)-2nd row and (c)-2nd row for image enhancement are helpful to capture the missed detection of pedestrians and vehicles in the original thermal images. This shown the potential of saliency maps applied to object detection in complex scenes. On the other hand, using only saliency maps to do the detection also might miss some potential objects, as shown in the 2nd row of (b) and (d). After the fusion of thermal images, these objects are captured successfully. The above results have shown the complementarity between thermal images and saliency maps, which proves the hypothesis that the fusion of saliency maps can improve the object detection accuracy of thermal images.

Furthermore, a quantitative comparison of the Average Precision (AP) of different networks is given in Table 1. The proposed model achieves the AP of 0.881 and 0.899 for pedestrians and vehicles that are superior over the other two.

Table 1. The average precision (AP) of the different models trained on different data.

Model training data	Pedestrian	Vehicle
Only thermal image	0.836	0.873
Only saliency map image	0.771	0.820
Fusion image	0.881	0.899

6. CONCLUSIONS

In this paper, we proposed an approach based on the fusion images of thermal images and the saliency maps to improve the performance of object detection (e.g. pedestrians and vehicles). Specifically, we get the boundary aware saliency maps by BASNet. As thermal-visible image pairs might not always be available, our research is focused on using only thermal images, eliminating the need for coupled visible images. Experimental results show that this method has a high potential for applications. The proposed technique can be used in multi-monitoring and emergency management tasks.

ACKNOWLEDGEMENTS

This work was supported in part by the National Natural Science Foundation of China under Grant: 41801342; the Natural Science Foundation of Jiangsu Province, China, under Grant: BK20170781; the Fundamental Research Funds for the Central Universities, NO.3082018NS2018030.

REFERENCES

Cheng, M., Mitra, N. J., Huang, X., Torr, P. H. S., Hu, S., 2015. Global contrast based salient region detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(3), 569-582, doi.org/10.1109/CVPR.2011.5995344.

Dollár, P., Tu, Z. W., Perona, P., Belongie, S., 2009. Integral channel features. *British Machine Vision Conference (BMVC)*, London, Britain, pp. 1-11.

Girshick, R., 2015. Fast R-CNN. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Santiago, Chile, pp. 1440-1448.

Girshick, R., Donahue, J., Darrell, T., Malik, J., 2014. Rich feature hierarchies for Aaccurate object detection and semantic segmentation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Columbus, USA, pp. 580-587.

He, S., Lau, R., Liu, W., Huang, Z., Yang, Q., 2015. SuperCNN: A super-pixel-wise convolutional neural network for salient object detection. *International Journal of Computer Vision*, 115(3), 330-344, doi.org/10.1007/s11263-015-0822-0.

Hou, Q.B., Cheng, M.M., Hu, X.W., Borji, A., Tu, Z.W., Torr, P., 2017. Deeply supervised salient object detection with short connections. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, USA, pp. 3203-3212

Hou, X.D., Zhang, L.Q., 2007. Saliency Detection: A spectral residual approach. *Proceedings of the 2007 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Minneapolis, USA.

Itti, L., Koch, C., Niebur, E., 1998. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1998, 20(11), 1254-1259.

Klein, D. A. and Frintrop, S., 2011. Center-surround divergence of feature statistics for salient object detection. *International Conference on Computer Vision (ICCV)*, Barcelona, Spain, pp. 2214-2219.

Li, C.Y., Song, D., Tong R.F., Tang M., 2019. Illumination-aware faster R-CNN for robust multispectral pedestrian detection. *Pattern Recognition*, 2019, 85, 161-171.

Li, M.H., Peng, L. B., Chen, Y.P., Huang, S. Q., Qin, F. Y. and Peng, Z. M., 2019. Mask sparse representation based on semantic features for thermal infrared target tracking, *Remote Sensing*, 11(17), 1967, doi.org/10.3390/rs11171967.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L., 2014. Microsoft coco:

Common objects in context. *European Conference on Computer Vision (ECCV)*, Zurich, Switzerland, pp. 740-755.

Olaf, R., Philipp, F., and Thomas, B., 2015. U-net: Convolutional networks for biomedical image segmentation, *International Conference on Medical image computing and computer-assisted intervention*, Munich, Germany, pp. 234-241.

Portmann, J., Lynen, S., Chli, M., Siegwart, R., 2014. People detection and tracking from aerial thermal views. *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, Hong Kong, China, pp.1794-1800.

Qin, X. B., Zhang, Z. C., Huang, C. Y., Gao, C., Dehghan, M. and Jagersand, M., 2019. BASNet: boundary-aware salient object detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, USA, pp. 7479-7489.

Redmon, J., Divvala, S., Girshick, R., Farhadi, A., 2016. You only look once: unified, real-time object detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, USA, pp. 779-788.

Redmon, J., Divvala, S., Girshick, R., Farhadi, A., 2017. YOLO9000: better, faster, stronger. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, USA, pp. 7263-7271.

Redmon, J., Farhadi, A., 2018. YOLOv3: an incremental improvement, [Online]. Available: https://arxiv.org/abs/1804.02767.

Ren, S. Q., He, K. M., Girshick, R., Sun J., 2017. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 39(6), 1137-1149.

Sun, Y., Yang, J. G., Li, M., An, W., 2019. Infrared small-faint target detection using non-i.i.d. mixture of Gaussians and flux density. *Remote Sensing*, 11(23), 2831, doi.org/10.3390/rs11232831.

Wada, K., 2016. Labelme: image polygonal annotation with python. https://github.com/wkentaro/labelme. Version: 4.2.9, Accessed: 10-02-2020.

Wang, P., Bai, X. Z., 2019. Thermal infrared pedestrian segmentation based on conditional GAN. *IEEE Transactions on Image Processing*, 28(12), 6007-6021.

Xu, D., Ouyang W. L., Ricci, E., Wang X.G., Sebe N., 2017. Learning cross-modal deep representations for robust pedestrian detection. *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Honolulu, USA: IEEE. 5363-5371.

Yann, L., Leon, B., Yoshua, B., Patrick, H., 1998. Gradient-based learning applied to document Recognition. *Proceedings of the IEEE*, 86(11), 2278-2324.

Yu, X. S., Liu, J. F., Tang, X. L., Huang, J. H., 2010. Estimating the pedestrian 3D motion indoor via hybrid tracking model. *ACTA AUTOMATICA SINICA*, 36(4), 610-615.

Zhang, F., Du, B., and Zhang, L., 2015. Saliency-guided unsupervised feature learning for scene classification. *IEEE Trans. Geosci. Remote Sens.*, 53(4), 2175-2184.

Zhang, X. W., Yang, Y. N., Yang, T., Zhang, X.G., Shao, D. P., 2010. Automatic visual-thermal image sequence registration based on co-motion. *ACTA AUTOMATICA SINICA*, 36(9): 1220-1231.