Chapter 29
Drowning Person Target Intelligent Recognition Method Based on Fusion of Visible Light and Infrared Thermal Imaging

Jianan Luo, Chunxu Li, and Jie Wen

Abstract Inland waterway, especially mountainous waterway channel, has the characteristics of rapid current, different width of river and large change of water level, which brings great risks to the navigation of ships. Once the persons fall into the water, it is difficult to search and rescue. This study aimed at developing a rapid drowning person recognition method, which establishes a deep learning architecture for infrared and visible image fusion. Compared with the traditional convolution network, the coding network is combined with convolution layer, fusion layer and dense block, in which the outputs of each layer are connected with each other, which can be used to obtain more useful features from the source image in the coding process. The target detection experiment of drowning personnel is carried out in the Lancang River. The results show that the method can accurately identify the target under the conditions of insufficient illumination and fast-moving speed, and the recognition rate is 90%.

29.1 Introduction

Every year, people drown all over the world. Tens of thousands of people die of drowning every year due to the accidental drowning of crew, tourists, capsizing and sinking of ships and so on. The main reason is that the current is turbulent and the water area is large, so it is difficult to find and locate the person falling into the water. With the upgrading of computing hardware and the optimization of artificial intelligence algorithms, image processing and detection have been applied to solve all kinds of problems, but the problem of drowning person detection still needs to be solved urgently.

Image fusion is an enhancement technology. Its purpose is to combine the images obtained by different types of sensors to generate images with stronger robustness or

J. Luo (✉) · C. Li · J. Wen
Intelligent Shipping Center, China, Waterborne Transport Research Institute, Beijing, China
e-mail: marinegis@foxmail.com

G. Liu and F. Cen (eds.), *Advances in Precision Instruments and Optical Engineering*,
Springer Proceedings in Physics 270, https://doi.org/10.1007/978-981-16-7258-3_29

richer information, so as to facilitate subsequent processing or help decision-making. Infrared and visible image fusion has advantages in many aspects.

First, their signals come from different forms, which provide different aspects of scene information, that is, the visible image captures the reflected light, while the infrared image captures the thermal radiation [1]. Therefore, this combination is more informative than the single-mode signal. Second, infrared and visible images show the inherent characteristics of almost all objects, which can be obtained by relatively simple equipment [2]. Finally, infrared image and visible image have complementary characteristics, so as to produce robust and informative fusion image. Visible images usually have high spatial resolution and considerable detail and light–dark contrast. Therefore, they are in line with human visual perception. However, these images are easily affected by bad conditions, such as insufficient lighting, fog and other bad weather. Infrared images describing the thermal radiation of objects can resist these interferences but usually have low resolution and poor texture. Due to the universality and complementarity of the images used, visible and infrared image fusion technology has a wider application field than other fusion technologies.

The fusion of visible and infrared images is of great significance for personnel detection, especially for people falling into the water. First of all, if only visible light images are used for detection, people are in the fast flowing and unclear river. In addition, the proportion of people exposed to the water when falling is very small, and the people falling into the river are almost integrated with the river, which is difficult to distinguish between the naked eye and the camera [3]. Even excellent detection algorithms are difficult to detect accurately, and the light conditions are good and fashionable, it cannot be detected at night or in heavy fog [4]. The infrared image can distinguish people from the background well. Because the human body has a higher temperature than the river water, the brightness of the human body reflected in the infrared image will be higher than the river water, so it is more prominent [5]. However, due to the low resolution and lack of texture features, the infrared image can only obtain rough contour information [6].

If there are high-temperature objects similar to the shape of the drowning person in the picture, it is easy to cause misjudgment and missing judgment, as shown in Fig. 29.1. If the visible and infrared images are fused, the image not only highlights the human body but also contains certain texture features, which will greatly improve the detection accuracy and recall.

29.2 Image Acquisition and Registration

29.2.1 Image Acquisition

The study uses a dual light camera that can obtain visible and infrared images at the same time. One side is an optical camera and the other is an infrared thermal imaging camera.

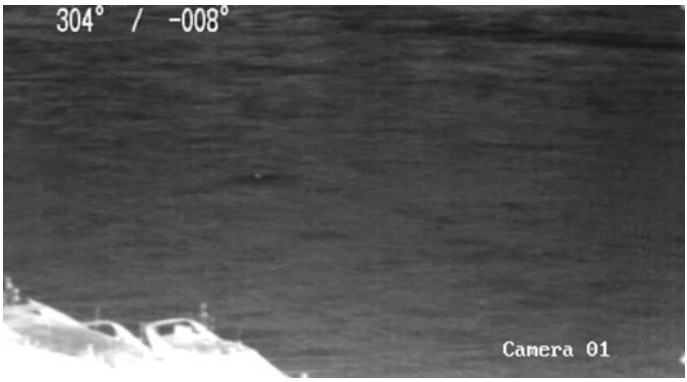

Fig. 29.1 Missing judgment of drowning personnel

29.2.2 Image Registration

Because infrared and visible images are obtained by different sensors, they are usually different in size, perspective and field of view [7]. The above dual light camera will also bring different viewing angles. However, successful image fusion requires strict geometric alignment of the fused image, so it is necessary to register the visible and infrared images before fusion. The registration of infrared image and visible image is a multi-mode registration problem.

For the registration problem here, the feature-based registration method is used. The feature-based method first extracts two groups of salient structures, then determines the correct correspondence between them, and estimates the spatial transformation accordingly, which is then used to align a given image pair.

The first step of feature-based method is to extract robust common features that can represent the original image. Edge information is one of the most commonly used choices in infrared and visible image registration, as shown in Fig. 29.2, because different registration methods can well preserve the size and direction of edge information. Edge mapping can be discretized into point sets. A popular strategy to solve the point matching problem includes two steps: calculating a set of assumed correspondences and then removing outliers through geometric constraints. By calculating feature descriptors at points, the matching between points with too large descriptor

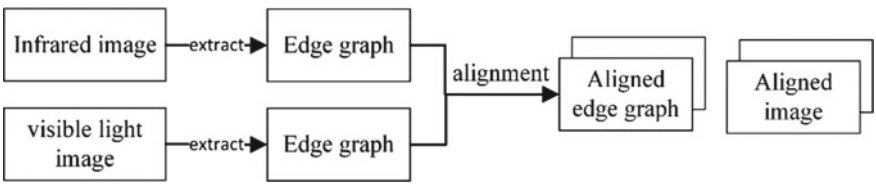

Fig. 29.2 Edge information features and image registration process

difference is eliminated, random sample consistency (RANSAC) is used to remove false matching from the assumed set, and the hypothesis verification method is used to obtain the minimum possible outlier without subset through resampling to estimate the given parameter model.

29.3 Image Fusion

29.3.1 Converged Network

A deep learning architecture for infrared and visible image fusion is adopted. Compared with the traditional convolution network, the coding network is combined with convolution layer, fusion layer and dense block, in which the output of each layer is connected with each other. Using this architecture, we can obtain more useful features from the source image in the coding process, select the appropriate fusion strategy to fuse the features, and finally reconstruct the fused image through the decoder.

As shown in Fig. 29.3, before fusion, the depth features of visible and infrared images are extracted, the first convolution layer extracts rough features, and then three convolution layers (the output of each layer is cascaded into the input of subsequent layers) form dense blocks. Such an architecture has two advantages. First, the size of the filter and the step of convolution operation are 3 respectively $\times 3$ and 1. Using this strategy, the input image can be any size; Second, dense blocks can retain depth features as much as possible in the coding network, and this operation can ensure that all salient features are used in the fusion strategy.

As shown in Fig. 29.4, L1 norm and soft-max operations are applied in the fusion layer. The fusion layer includes a plurality of convolution layers (3×3), the output of the fusion layer will be the input of the convolution layer. This simple and effective architecture is used to reconstruct the final fused image.

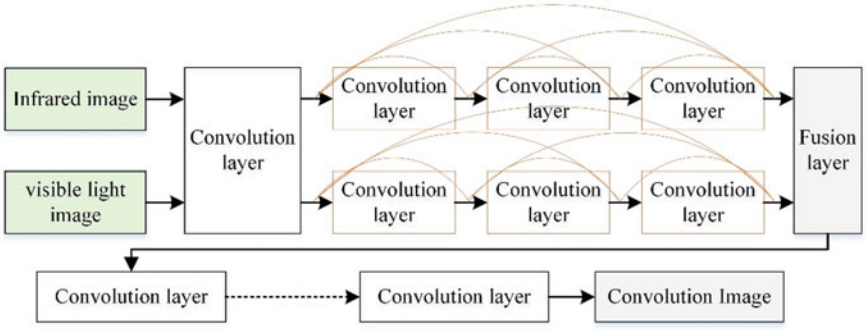

Fig. 29.3 Fusion network structure

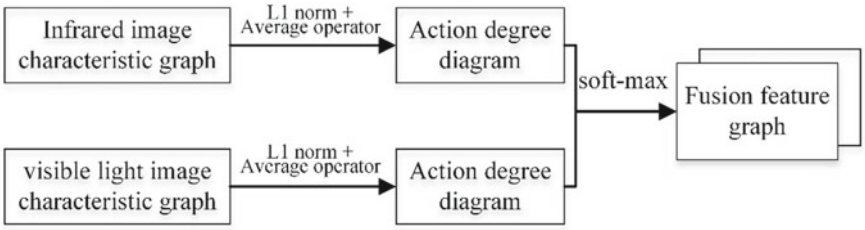

Fig. 29.4 L1 norm and softmax operation are applied in the fusion layer

29.3.2 Loss Function

The loss function of fusion network is composed of pixel loss function L_P and structural similarity loss function L_{ssim} weighting results in:

$$L_p = \|O - I\|^2 \tag{29.1}$$

$$L_{ssim} = 1 - SSIM(O,\ I) \tag{29.2}$$

$$L_{fus} = \lambda L_{ssim} + L_p \tag{29.3}$$

where O and I represent an output image and an input image, respectively. L_p is the Euclidean distance between output O and input I. L_{ssim} represents the structural similarity, which represents the structural similarity of two images. This index is mainly composed of three parts: correlation, brightness loss and contrast distortion. The product of the three components is the evaluation result of the fused image. Since there are three orders of magnitude differences between pixel loss and L_{ssim} loss, in the training phase, the λ set to 1000.

29.4 Detection of Personnel Falling into the Water

29.4.1 Detection Network

The convolutional neural network CNN is used to recognize the target of the drowning person. The central idea of the detection network is to divide the picture into S × S areas. If the center of an object falls on a cell, the cell is responsible for predicting the object. Each cell needs to predict multiple bounding box values, predict a confidence level for each bounding box, and then conduct prediction analysis in units of each cell.

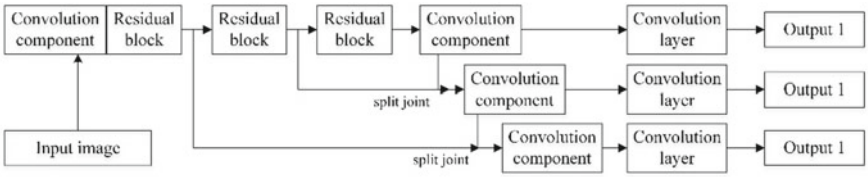

Fig. 29.5 Darknet-53 backbone network

The backbone network adopts the modified darknet-53, as shown in Fig. 29.5. This network has high classification accuracy, fast calculation speed and few network layers. The full connection layer is removed. The network here is a full convolution network, which uses a large number of residual layer hopping connections. In order to reduce the negative gradient effect caused by pooling, the pooling layer is abandoned and the step size of the convolution layer is used to realize downsampling. In this network structure, the convolution with step size of 2 is used for down sampling.

The network outputs three feature maps of different scales, draws lessons from FPN, and uses multi-scale to detect targets of different sizes. The finer the unit, the finer the object can be detected.

Before model training, it is first necessary to make a dataset of fusion images, capture visible and infrared images with a dual light camera, obtain the fusion images through the above registration and fusion process, label the drowning personnel, make a dataset in the format required for training, and select the pretraining model for training, The algorithm model that can identify the drowning person in the visible and infrared fusion image is obtained. Then evaluate the accuracy of the model and optimize it from the aspects of data set and algorithm, so that it can achieve a better recognition effect.

29.4.2 Loss Function

The loss function of the detection model is divided into three parts, L_{box} brought by bounding box, L_{obj} caused by confidence, error L_{cls} brought by category:

$$L_{box} = \lambda_{coord} \sum_{i=0}^{S^2} \sum_{j=0}^{B} 1_{i,j}^{obj} (2 - w_i \times h_i) \Big[\left(x_i - \hat{x}_i\right)^2 + \left(y_i - \hat{y}_i\right)^2 + \left(w_i - \hat{w}_i\right)^2 + \left(h_i - \hat{h}_i\right)^2 \Big] \tag{29.4}$$

$$L_{cls} = \lambda_{class} \sum_{i=0}^{S^2} \sum_{j=0}^{B} 1_{i,j}^{obj} \sum_{c \in classes} p_i(c) log\left(\hat{p}_i(c)\right) \tag{29.5}$$

$$L_{obj} = \lambda_{nobj} \sum_{i=0}^{S^2} \sum_{j=0}^{B} 1_{i,j}^{nobj} \left(c_i - \hat{c}_i\right)^2 + \lambda_{obj} \sum_{i=0}^{S^2} \sum_{j=0}^{B} 1_{i,j}^{obj} \left(c_i - \hat{c}_i\right)^2 \tag{29.6}$$

The detection loss function is the sum of the above three errors:

$$L_{dec} = L_{box} + L_{cls} + L_{obj} \tag{29.7}$$

$$L = L_{fus} + L_{dec} \tag{29.8}$$

29.5 Detection Fusion Reverse Guidance

The purpose of common visible infrared image fusion technology is to make the fused image contain as much information of two kinds of images as possible, neither lose the contrast information in the infrared image nor the texture information in the visible image, or make the fused image more in line with the human visual system, Therefore, the loss function of the initial fusion process is defined as the weighted sum of the pixel loss function and the structural similarity loss function.

The focus of this system is to accurately detect the person falling into the water. The result of image fusion is only an intermediate process. Whether it is image fusion or detection process, its optimization should take accurate detection as the ultimate goal. In order to achieve this ultimate goal, the training of image fusion should be modified so that the loss function in the detection process can guide the fusion, and the final detection results will be optimized in the fusion stage.

As shown in Fig. 29.6, first mark the person falling into the water on the registered visible or infrared image. Since the image has been registered and aligned, and the target position after fusion remains unchanged, the mark can be copied to the fusion image as the ground truth. After the fusion image passes through the detection network, the predicted boundary box, classification and confidence are obtained, and the detection error is calculated by comparison with the mark, i.e. L_{dec}, this loss

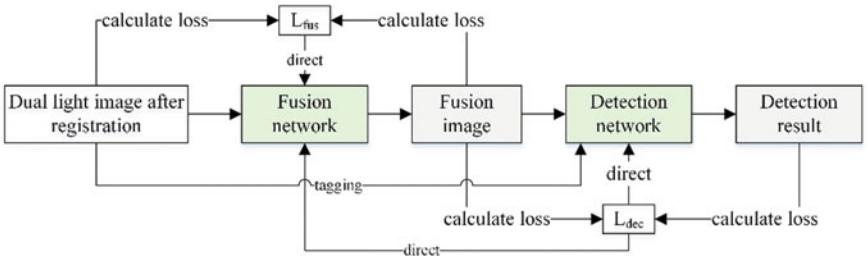

Fig. 29.6 Dual registration image fusion detection

function is used not only to evaluate and optimize the detection network but also to evaluate and optimize the fusion network. It is equivalent to the loss function of the fusion network, and the following corrections are made:

$$L = L_{fus} + L_{dec} \tag{29.9}$$

Furthermore, we developed a waterway operational monitoring system based on this study and demonstrated its application in China's inland waterway, as shown in the figure below. The test shows that the recognition accuracy of the system at night is more than 95% (Figs. 29.7, 29.8, 29.9 and 29.10).

Fig. 29.7 Dynamic identification effect of crew in daytime

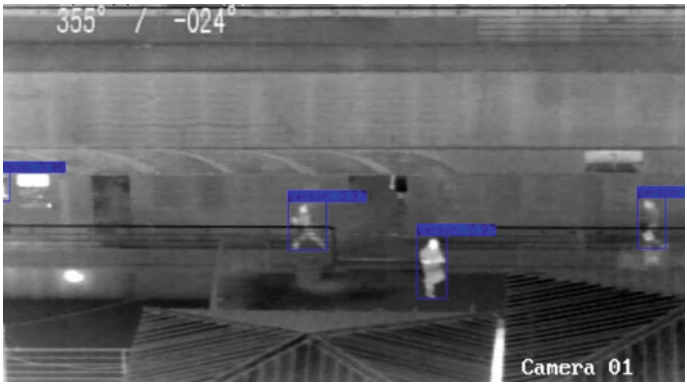

Fig. 29.8 Identification effect of crew at night

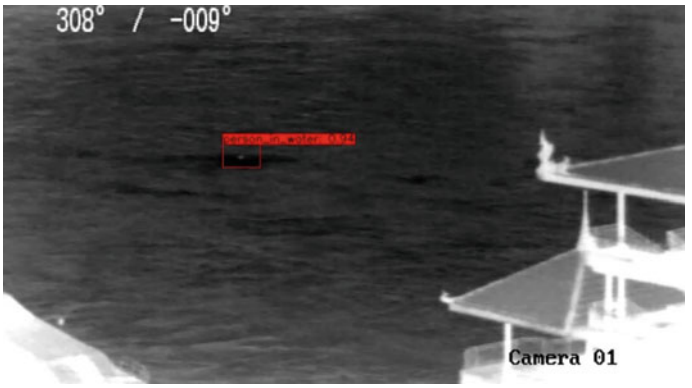

Fig. 29.9 Detection results of people falling into the water under thermal imaging mode (depression angle 9°)

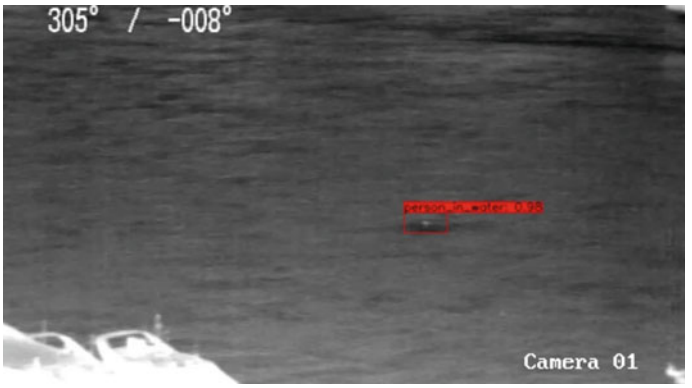

Fig. 29.10 Detection results of people falling into the water under thermal imaging mode (depression angle 8°)

29.6 Conclusion

In this study, a drowning person detection method based on visible light and thermal imaging data fusion is proposed. The method includes image acquisition, image registration, image fusion and target detection. It is a complete, feasible and practical system, which can be used to detect drowning persons for subsequent positioning and rescue.

Acknowledgements This work was supported by China Waterborne Transport Research Institute Fundamental Research Funds under Grant 182102-2021.

References

1. W. Ronggui, W. Jing, Y. Juan, Feature pyramid random fusion network for visible-infrared modality person re-identification. Opto-Electron. Eng. Sichuan **47**, 190669-1–190669-12 (2020)
2. Q. Yonsheng, S. Guobing, W. Yuwu, Research on night environment image enhancement algorithm based on infrared and visible light image fusion. J. Harbin Univ. Commer. (Nat. Sci. Ed.). Haerbin **37**, 422–427 (2021)
3. D. Guipeng, T. Gang, L. Chunying et al., Infrared and visible images fusion based on non-subsampled contourlet transform and guided filter. Acta Armamentarii Beijing (in press)
4. G. Jiamin, L. Aiping, M. Doudou et al., Infrared and visible image fusion combining neighborhood features with IDCSCM. Laser Infrared **50**, 889–896 (2020)
5. Z. Xiaopeng, Feature pyramid random fusion network for visible-infrared modality person re-identification. Opto-Electron. Eng. Sichuan **47**, 190669-1–190669-12 (2020)
6. Y. Yongwu, Application of infrared image and neural network in ship target recognition. Ship Sci. Technol. Beijing **43**, 175–177 (2021)
7. S. Jianhui, Z. Hang, L. Jianju, Infrared and visible light image fusion based on DPN deep learning network. J. Shenyang Ligong Univ. Shenyang **39**, 2023–2027 (2020)