Face Detection on Thermal Infrared Images Combined with Visible Images

Yujia Chen[1], Liqing Wang[2], and Guangda Xu[3(✉)]

[1] School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an, China
jerryniu520126@foxmail.com
[2] Infrared Division, Wuhan Huazhong Numerical Control Co., Ltd., Wuhan, China
[3] National Numerical Control System Engineering Research Center, Huazhong University of Science and Technology, Wuhan, China
xu_guangda@hust.edu.cn

Abstract. Due to COVID-19, intelligent thermal imagers are widely used all over the world. Since intelligent thermal imagers usually require real-time temperature measurement, it is significant to find a method to quickly and accurately detect human faces in thermal infrared images. This paper mainly proposes two different methods. One is to use image processing methods and face detected from visible images to determine the position of the face in the infrared image, while the other is to use target detection algorithms on infrared images, including YOLOv3 and Faster R-CNN. This paper uses the two methods on a self-collected dataset containing 944 pairs of visible and infrared images and observes the robustness of methods by adding random noise to images. Experiments show that the first one has much lower latency and the latter one has higher accuracy in both cases.

Keywords: Face detection · Thermal infrared image · YOLOv3 · Faster R-CNN · Intelligent thermal imagers

1 Introduction

Visible images refer to images formed by visible light (light with a wavelength between 390 nm–780 nm). The photos taken by ordinary cameras are visible images. In recent years, target detection, especially face detection for visible images, is developing rapidly and many effective algorithms can be used. For example, it is a good choice to use R-CNN (Region-based Convolutional Neural Networks) [5] and its derivative algorithms, including Fast R-CNN [4], Faster R-CNN [14], Mask R-CNN [6] and R-FCN [3]. Different versions of YOLO (You only look once) are also effective, including YOLOv1 [11], YOLOv2 [12], YOLOv3 [13], YOLOv3 [1] and PP-YOLO [10].

Because of COVID-19, intelligent thermal imagers are widely used all over the world. To prevent interference caused by the ambient temperature, the intelligent thermal imager needs to narrow down the temperature measurement area.

L. Pan et al. (Eds.): BIC-TA 2021, CCIS 1566, pp. 346–357, 2022.
https://doi.org/10.1007/978-981-19-1253-5_26

Given that the skin of the human face is exposed to the air, the human face is considered a suitable temperature measurement area. Therefore, an intelligent thermal imager usually needs to find the position of the human face in the infrared image as the temperature measurement area [16]. Face detection in thermal infrared images becomes an important issue.

In recent years, there are some papers on face detection in thermal infrared images [7–9,15]. Most previous papers used high-resolution single-person thermal infrared images with little noise in the environment. However, in actual application scenarios of intelligent thermal imagers, thermal infrared images may have low resolution, many people, or much environmental noise. The facial features in the thermal infrared image are probably not visible due to low resolution, which means they may perform badly in a real situation.

In addition, since intelligent thermal imagers usually require real-time temperature measurement, they usually do not use target detection algorithms directly on infrared images because their processing speed is not fast enough. Instead, many intelligent thermal imagers copy the faces detected in visible images to the corresponding infrared images [16]. However, the results of this method are not very accurate. For example, Fig. 1 is a screenshot of HY-2005B intelligent thermal imager, produced by Wuhan Huazhong Numerical Control Co., Ltd. [17].

Fig. 1. A screenshot of an intelligent thermal imager

Figure 1 shows that faces detected in visible images are quite precise, while faces detected in infrared images have noticeable offsets. The main reason is that these two different types of images are captured by two different cameras and there is a small distance between them (shown as Fig. 2), which leads to different optical axes of the two cameras. This means that even the same pixel in the infrared image and the visible image corresponds to a different position.

In conclusion, it is significant to find a method to quickly and accurately detect human faces in thermal infrared images with corresponding visible images.

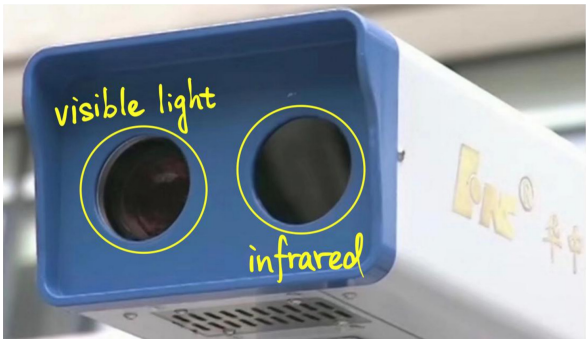

Fig. 2. The infrared camera and the visible light camera

2 Principle of Image Calibration

Given that the distance between the two cameras is small, it can be assumed that the mapping between infrared images and visible images is a rigid transformation. In other words, the pixel (x, y) in the visible image and the pixel (x', y') in the infrared image satisfy the following relationship (shown as Eq. (1)).

$$\begin{cases} x' = a_{11}x + a_{12}y + a_{13} \\ y' = a_{21}x + a_{22}y + a_{23} \end{cases} \tag{1}$$

where $a_{ij}(i = 1, 2; j = 1, 2, 3)$ are all constants.

The matrixed equation is as follows

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix} \cdot \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \mathbf{A} \cdot \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \tag{2}$$

If the coordinates of the three corresponding point pairs are known and they are $(x_k, y_k)(x'_k, y'_k)(k = 1, 2, 3)$, then the parameters $a_{ij}(i = 1, 2; j = 1, 2, 3)$ can be determined by the following equation:

$$\begin{cases} x'_k = a_{11}x_k + a_{12}y_k + a_{13} & (k = 1, 2, 3) \\ y'_k = a_{21}x_k + a_{22}y_k + a_{23} & (k = 1, 2, 3) \end{cases} \tag{3}$$

This method is simple and easy to implement. By using this method, people can determine the position of human faces in the infrared image through the face box in the corresponding visible image.

However, it should be noticed that this method ignores the distance between the object and the camera, which will affect the calibration. For example, if all the calibration point pairs are obtained at a distance of 5 m from the camera, the calibration of the object at 2 m will have an offset. This paper will analyze the causes of offset in detail and propose ways to reduce it.

3 Optical Analysis for Calibration Offset

This section tries to explain the reasons why the calibration offset occurs by performing an optical analysis.

3.1 Optical Model Derivation

The optical model of the intelligent thermal imager is shown in Fig. 3. $\overline{BG}$ represents the optical axis of the visible camera, while $\overline{CT}$ represents the optical axis of the infrared camera. $\overline{FG}$ and $\overline{ST}$ are the light screens of the visible camera and infrared camera respectively. Note that the light screens are on the same plane. The point P represents the object being photographed.

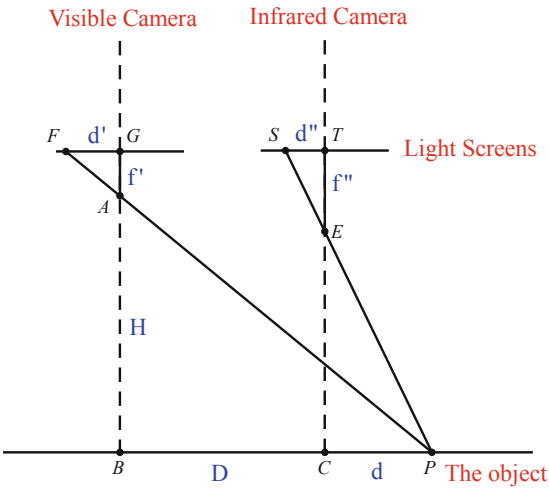

Fig. 3. The optical representation of an intelligent thermal imager

The geometric similarity tells that

$$\frac{FG}{BP} = \frac{AG}{AB}$$

$$\frac{ST}{CP} = \frac{ET}{CE}$$

Namely,

$$\frac{d'}{D+d} = \frac{f'}{H} \tag{4}$$

$$\frac{d''}{d} = \frac{f''}{H+f'-f''} \tag{5}$$

Since $H \gg f', f''$ and $f' \approx f''$, $H \gg f' - f''$. Thus, the Eq. (5) can be simplified as

$$\frac{d''}{d} = \frac{f''}{H} \tag{6}$$

By combining the Eqs. (4) and (6), the following relationship can be obtained

$$d'' = d' \cdot \frac{f''}{f'} - \frac{Df''}{H} \tag{7}$$

Let $k = \frac{f''}{f'}$ and $C = \frac{Df''}{H}$. Then the Eq. (7) becomes

$$d'' = d' \cdot k - C \tag{8}$$

Suppose that at a fixed distance H_0, take any two calibration points. The distances between these two points and point C are d_1 and d_2 respectively. d_1' and d_1'' correspond to d_1, while d_2' and d_2'' correspond to d_2. Then, the following linear equations can be obtained

$$\begin{cases} d_1'' = d_1' \cdot k - C \\ d_2'' = d_2' \cdot k - C \end{cases} \tag{9}$$

By solving the Eq. (9), k and C can be determined.

The calibration method introduced in Sect. 2 regards the distance between the object and the two cameras H as a constant H_0. In other words, This calibration method only uses constants k and C for calibration, regardless of the distance between the object and the camera. However, H is changing, which causes C is also changing. That is the reason why the calibration has an offset.

3.2 Analysis for Calibration Offset

Assume that the calibration is performed at H_0. Then, the calibration offset is $\frac{Df''}{H} - \frac{Df''}{H_0}$. When $H > H_0$ (that is, when the actual object is farther from the camera), $\frac{Df''}{H} - \frac{Df''}{H_0} < 0$ because D and f'' keep unchanged. This means that d'' will be smaller than it should be, if we still use k and C obtained by H_0 for calibration in this case. Similarly, when $H < H_0$ (that is, the actual object is closer to the camera), d'' will be larger than it should be.

However, it should be noticed that $\frac{1}{H}$ is an inverse function. Hence, the calibration offset at $H < H_0$ is much larger than the calibration offset at $H > H_0$. For example, suppose $H_0 = 5$ meters. On the one hand, when $H = 1$ meter, the calibration offset is

$$\frac{Df''}{H} - \frac{Df''}{H_0} = \frac{4}{5} Df''$$

On the other hand, when $H = 9$ meters, the calibration offset is

$$\frac{Df''}{H} - \frac{Df''}{H_0} = -\frac{4}{45}Df''$$

In this case, $\frac{4}{5}Df'' > \frac{4}{45}Df''$.

3.3 Methods to Reduce Calibration Offset

The first two methods try to reduce the calibration offset on the physical level, while the latter two methods try to rectify the calibration results from the perspective of image processing.

Method 1: Using the Exact H Instead of H_0. This method can even eliminate the calibration offset. However, it should be noted that this method requires real-time measurement of the distance between each object and the camera, which needs very professional equipment and incurs huge costs. Therefore, this method is theoretically usable but not practically affordable.

Method 2: Decreasing the Distance Between the Two Cameras. Since the calibration offset $\frac{Df''}{H} - \frac{Df''}{H_0}$ is linear to the distance between the two cameras D, decreasing it is another choice. However, it should be noticed that the ideal situation $D = 0$ cannot be implemented physically because the visible camera and the infrared camera are two different cameras.

Method 3: Looking for the Largest Connected White Area. This method mainly has 4 steps (shown in Fig. 4). The largest connected white area is the precise face box as required.

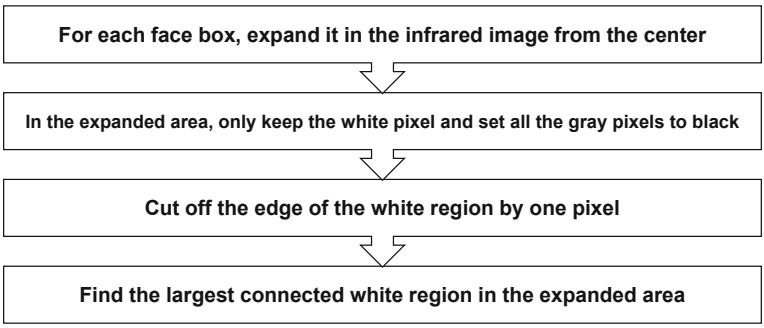

Fig. 4. A flowchart of the procedure of Method 3

Take Fig. 5 as an example. The red rectangle in Fig. 5 shows the face box obtained by using the calibration method above.

First, expand the red rectangle by half of its side length to obtain the green rectangle in Fig. 5.

Fig. 5. An example for Method 3 (Color figure online)

Second, set all the pixels with a gray-scale value between 221 and 255 to 1 (i.e. white) and setting the other pixels to 0 (i.e. black).

Next, traverse all the white pixels and determine whether the four pixels up, down, left, and right are all 1 (i.e. white). Save the positions of the pixels that do not meet this condition and set them to 0 (i.e. black).

Finally, use the seed-filling algorithm or the two-pass algorithm to find all the connected white regions, which are both traditional and classical algorithms to do that. Then compute the area of all connected regions and find the largest one, which is the blue rectangle in Fig. 5.

Method 4: Using Target Detection Algorithms. Method 3 uses classic image processing methods to detect faces. Thus, combining some more advanced algorithms seems to be a good option, especially some mature and effective target detection algorithms, such as Faster-RCNN [14] or YOLOv3 [13].

The first step of Method 4 is the same as the one of Method 3, which is expanding each face box from the center. However, Method 4 directly uses target detection algorithms to detect a human face in the expanded area, while Method 3 uses 3 more steps to process the infrared image.

4 Experiments and Results

4.1 Experimental Setup

Data Collection. Since there are few pairs of infrared and visible images in the public data set containing clear human faces, this paper uses a self-collected data set containing 944 pairs of infrared and visible images. Some pairs of images are shown in Fig. 6. All the infrared images are grayscale images consisting of 256 * 320 pixels, while all the visible images are RGB images consisting of 1024

* 1280 pixels. The faces in each image are labeled with rectangular boxes as ground truths. Labels are stored in XML files.

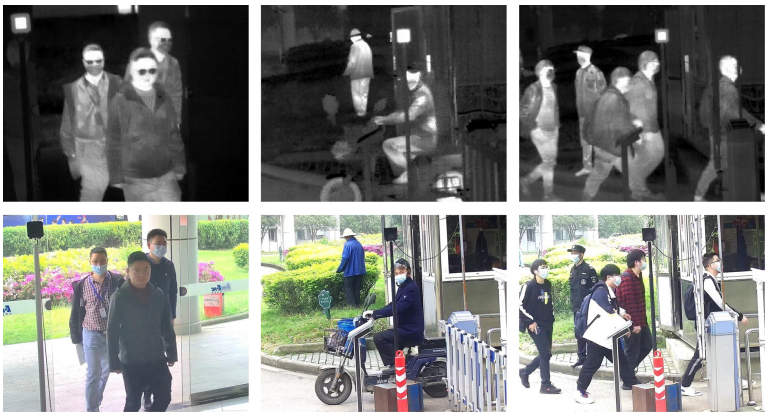

Fig. 6. A few examples of the data set

For data augmentation, this paper performs a horizontal flip for all image pairs. Besides, this paper standardizes all the images before training, evaluation, and final testing.

This paper shuffles all preprocessed image pairs and extracts 70% of them as the training set, 15% as the validation set, and the remaining 15% as the test set.

Performance Measurement. This paper mainly uses two parameters to measure the performance of methods: average precision (AP) when the threshold is intersection over union (IoU) = 0.5 and frames per second (FPS) in the following experimental environment

GPU: Tesla V100. Video Mem: 16 GB
CPU: 4 Cores. RAM: 32 GB. Disk: 100 GB

For each method, the AP and FPS of the optimal model on the test set will be collected as its final performance.

Optimization and Evaluation. This paper uses a trained PP-YOLO model to detect faces in visible images, whose FPS is about 50 in the above environment [10].

For Method 4, this paper uses two algorithms, YOLOv3 [13] and Faster R-CNN [14], to detect faces in infrared images. The optimization and evaluation strategy is shown in Table 1, where LR is the abbreviation of learning rate.

This paper mainly compares the performance of the following five methods.

Table 1. Optimization and evaluation strategy

Algorithms	Backbone	LR	Epochs	LR decay epochs	Optimizer
YOLOv3	DarkNet53	0.000125	50	[30, 40]	Momentum
Faster R-CNN	ResNet50	0.0025	12	[10, 11]	Momentum

1. PP-YOLO + Image Processing (Method 3)
2. PP-YOLO + YOLOv3 (Method 4)
3. YOLOv3 only
4. PP-YOLO + Faster R-CNN (Method 4)
5. Faster R-CNN only

The optimization and evaluation strategy of PP-YOLO + YOLOv3 is the same as YOLOv3 only. Similarly, the optimization and evaluation strategy of PP-YOLO + Faster R-CNN is the same as Faster R-CNN only.

Random Noise. In order to compare the robustness of methods, this paper conducts experiments in two situations: with random noise and without random noise.

The random noise is added by randomly scaling the brightness and contrast of images. The probability of whether randomly scaling contrast or brightness is 0.5. When scaling is required according to the random result, take a random value from $[0.5, 1.5]$ as the scaling factor s_b, s_c and distort the image according to the following formulas.

$$\text{Brightness} : \mathbf{Image}(i,j) \Longrightarrow s_b \cdot \mathbf{Image}(i,j)$$

$$\text{Contrast} : \mathbf{Image}(i,j) \Longrightarrow s_c \cdot \mathbf{Image}(i,j) + (1 - s_c) \cdot \overline{\mathbf{Image}}$$

where $\mathbf{Image}(i,j)$ refers to each pixel and $\overline{\mathbf{Image}}$ is the mean of the image. Namely,

$$\overline{\mathbf{Image}} = \frac{1}{M \cdot N} \sum_{i=1}^{M} \sum_{j=1}^{N} \mathbf{Image}(i,j)$$

4.2 Results

The experimental results are shown in Table 2.

Experiments show that Method 3 has a much higher FPS than the other methods, which means that Method 3 can quickly find the positions of human faces in infrared images. However, the AP of Method 3 is much worse than Method 4's AP. That is mainly because Method 3 has weaker robustness than that of Method 4. For example, if a person wears a mask, Method 3 tends to

Table 2. The experimental results

Algorithms	Without noise		With noise	
	AP (%)	FPS (Hz)	AP (%)	FPS (Hz)
PP-YOLO + Image Processing	58.13	42	49.64	36
PP-YOLO + YOLOv3	91.03	11	91.05	10
YOLOv3 only	90.33	16	90.27	14
PP-YOLO + Faster R-CNN	91.60	10	91.62	9
Faster R-CNN only	90.80	13	90.83	13

output the part without the mask, while the ground truth often includes the mask.

However, to an intelligent thermal imager, only the face area for temperature measurement needs to be determined. Those occluded parts of the face (such as the area covered by a mask) will not be used for parking spaces. Therefore, although Method 3 cannot detect a complete face in some cases, it can still be used in this scenario.

Therefore, if high precision is required, Method 4 is better; if low latency is required, Method 3 is better.

5 Conclusion

This paper mainly proposes two methods to solve the problem of face detection in infrared images. One (called Method 3) is to uses image processing methods and face detected from visible images to determine the position of the face in the infrared image, while the other (called Method 4) is to uses target detection algorithms on infrared images, including YOLOv3 and Faster R-CNN.

To be specific, Method 3 mainly consists of 4 steps:

Step 1: For each face box, expand it in the infrared image from the center
Step 2: In the expanded area, only keep the white pixel and set all the gray pixels to black
Step 3: Cut off the edge of the white region by one pixel
Step 4: Find the largest connected white region in the expanded area

The first step of Method 4 is the same as the one of Method 3, which is expanding each face box from the center. However, Method 4 directly uses target detection algorithms to detect a human face in the expanded area, while Method 3 uses 3 more steps to process the infrared image.

Experiments show that Method 3 can quickly find the positions of human faces in infrared images. However, Method 3 has weaker robustness than that of Method 4, which is the main reason why the performance of Method 3 is worse

than Method 4's performance. Therefore, if high precision is required, Method 4 is better; if low latency is required, Method 3 is better.

Besides, the methods proposed in this paper still have room for improvement in the correlation between visible and infrared images. This paper uses the calibration method proposed in Sect. 2 to associate visible and infrared images. In the future, other association methods are also worth trying (such as Cross-Modal CNN [2]).

References

1. Bochkovskiy, A., Wang, C.Y., Liao, H.Y.M.: YOLOV4: optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934 (2020)
2. Cangea, C., Veličković, P., Liò, P.: XFlow: cross-modal deep neural networks for audiovisual classification. IEEE Trans. Neural Netw. Learn. Syst. **31**(9), 3711–3720 (2019)
3. Dai, J., Li, Y., He, K., Sun, J.: R-FCN: object detection via region-based fully convolutional networks. In: Advances in Neural Information Processing Systems, pp. 379–387 (2016)
4. Girshick, R.: Fast R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1440–1448 (2015)
5. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 580–587 (2014)
6. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2961–2969 (2017)
7. Kopaczka, M., Nestler, J., Merhof, D.: Face detection in thermal infrared images: a comparison of algorithm- and machine-learning-based approaches. In: Blanc-Talon, J., Penne, R., Philips, W., Popescu, D., Scheunders, P. (eds.) ACIVS 2017. LNCS, vol. 10617, pp. 518–529. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-70353-4_44
8. Kopaczka, M., Schock, J., Nestler, J., Kielholz, K., Merhof, D.: A combined modular system for face detection, head pose estimation, face tracking and emotion recognition in thermal infrared images. In: 2018 IEEE International Conference on Imaging Systems and Techniques (IST), pp. 1–6. IEEE (2018)
9. Kowalski, M., Grudzień, A., Ciurapiński, W.: Detection of human faces in thermal infrared images. Metrol. Measure. Syst. **28**, 307–321 (2021)
10. Long, X., et al.: PP-YOLO: an effective and efficient implementation of object detector. arXiv preprint arXiv:2007.12099 (2020)
11. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 779–788 (2016)
12. Redmon, J., Farhadi, A.: YOLO9000: better, faster, stronger. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7263–7271 (2017)
13. Redmon, J., Farhadi, A.: YOLOV3: an incremental improvement. arXiv preprint arXiv:1804.02767 (2018)
14. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. Adv. Neural. Inf. Process. Syst. **28**, 91–99 (2015)

15. Ribeiro, R.F., Fernandes, J.M., Neves, A.J.: Face detection on infrared thermal image. SIGNAL 2017 Editors, p. 45 (2017)
16. Wang, H., Bai, C., Wang, J., Yuan, Z., Li, J., Gao, W.: The application of intelligent thermal imagers in response to COVID-19. China Metrol. **5**, 119–132 (2020)
17. Wuhan Huazhong Numerical Control Co., Ltd. (HCNC): Infrared AI thermal imagimg camera thermometer sensor face recognition body temperature scanner with CCTV dual thermal imager. https://www.alibaba.com/product-detail/Infrared-AI-Thermal-Imagimg-Camera-Thermometer_62490883868.html?spm=a2700.galleryofferlist.normal_offer.d_title.706950c9Bn2BwU. Accessed 04 May 2021