Displays 81 (2024) 102613

Contents lists available at ScienceDirect

Displays

journal homepage: www.elsevier.com/locate/displa

Physiological characteristics inspired hidden human object detection model ☆

Menghan Hu [a,1], Lejing Zhang [a,1], Bailiang Zhao [b], Yunlu Wang [a], Qingli Li [a], Lianghui Ding [c,*], Yuan Cao [d]

[a] Shanghai Key Laboratory of Multidimensional Information Processing, School of Communication and Electronic Engineering, East China Normal University, Shanghai 200241, China
[b] School of Health Science and Engineering, University of Shanghai for Science and Technology, Shanghai 200241, China
[c] Institute of Image Communication and Network Engineering, Shanghai Jiao Tong University, Shanghai 200240, China
[d] Naval Research Academy, Beijing 100161, China

ARTICLE INFO

MSC:
41A05
41A10
65D05
65D17

Keywords:
Occluded human detection
Proposal generation
Physiological inspired model
Improved selective search method

ABSTRACT

The current target detection algorithms provide the unsatisfactory performance on the task of detecting hidden human targets. In this study, we put forward the physiological characteristics inspired hidden human object detection model considering the spatio-temporal physiological features and their interdependent relationships. The experimental results of homemade hidden human object dataset demonstrate that the proposed model generates the detection accuracy of 64%, 44%, and 54% for indoor scene, outdoor scene, and overall dataset, respectively, outperforming the YOLO v4 models and the models based on HOG, LBP, and Haar features, with at least 22% promotion in detection accuracy. The ablation experiments indicate the effectiveness of each module of the method. In the future, the proposed model or the corresponding modeling idea has the potential to be applied to military rescue, public security investigation and other fields. Once the paper is accepted, we will make the homemade dataset publicly available.

1. Introduction

For human target detection, the current target detection algorithms such as the YOLO series models can give near-perfect detection results if there are complete or partially complete human target features in the image [1–5]. When human limbs are obscured by objects or humans hide themselves on purpose behind shelters such as walls, chairs, curtains and wood panels, the general object detection models and human body detectors or detection models tend to fail due to the absence of image features that are closely related to the human body. The phenomenon of human occlusion is very common in life [6–8]. The hidden human target detection task is somewhat similar to the small object detection task, but there is difference: the main difficulty in small target detection task is that the detected target is too small [9]; the difficulty in hidden human target detection task is not only that the detected target is too small, but also that the target features are missing. Due to the high complexity and strong pertinence, if the small target detection algorithms are applied for the hidden human target detection task, the model training cannot converge. Although the image features related to human body are missing due to occlusion, the inherent physiological features of human body can still be detected through video. Therefore, in hidden human target detection task, we innovatively introduce the physiological information to extend the temporal dimension to compensate for the lack of spatial dimension. In this paper, the Imaging Photo Plethysmo Graphy (IPPG) [10,11] is employed to recover human physiological signals from video, and the extracted signals are then aggregated into the hidden human target detection algorithm to detect hidden human targets from the background, thus overcoming the disadvantages of existing target detection algorithms.

Up to now, this is the first time that IPPG has been applied for hidden human target detection task [12]. The IPPG research is currently focused on how to measure human physiological signals consistently and accurately. Hu et al. [13] used the neural network to extract spatio-temporal information from videos, and performed information fusion to improve the noise resistance of the heart rate detection model. Comas et al. [14] implemented the recursive time-series UNet network for heart rate detection in the near-infrared surveillance camera scene, and its performance was verified to be better than existing hand-crafted models and end-to-end networks. Pai et al. [15] developed the

☆ This paper was recommended for publication by Prof Guangtao Zhai.

* Corresponding author.
E-mail address: lhding@sjtu.edu.cn (L. Ding).

[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.displa.2023.102613
Received 8 June 2023; Received in revised form 28 October 2023; Accepted 5 December 2023
Available online 15 December 2023

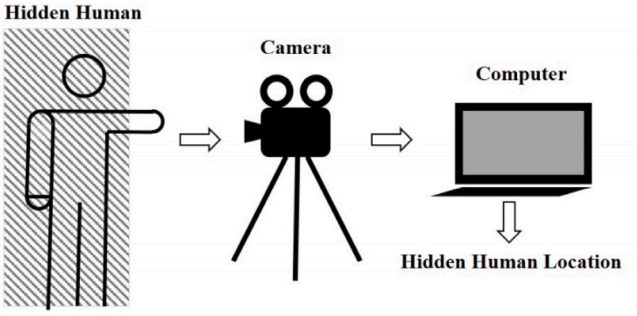

Fig. 1. Description of typical application scenario. The human target hides behind the shelter, and the video data is captured by RGB camera, and then the human body location is computed by the proposed method.

heart rate estimation algorithm in frequency domain to reduce motion-induced artifacts, and validated the algorithm on the self-built dataset. The research progresses of IPPG testify the ripeness and completeness in measuring physiological signals with the employment of the RGB camera.

For human target detection, the traditional operators such as SIFT [16], Haar [17], and HOG [18] have been designed to characterize the human body. A series of weak classifiers are subsequently used to detect the human targets. With the improvement of computing power and the development of neural network learning, the end-to-end models such as YOLO series networks are widely used in object detection tasks because of its ability to balance detection accuracy and speed [19]. Compared to RCNN series models, the YOLO series networks convert the object detection problem into single-stage regression problem. Currently, the above methods have not been used for the hidden human target task. In addition, it remains to be verified whether the above object detection algorithms can be effectively applied to hidden human target detection task.

The human targets will respond specifically in special imaging systems such as the hyperspectral imaging system [20], thermal imaging systems [21], FMCW radar [22]. However, these imaging devices are not readily available everywhere. To meet the needs of public security, setting up surveillance cameras is the regular way of many units or national departments. Therefore, it is critical to use available and accessible equipment to detect human targets in hidden situations. For example, in military security and public security investigations, detecting obscured human bodies allows staff to search and find suspects faster, which in turn improves the efficiency of government agencies and lessens needless human labor. Fig. 1 is the hypothetical application scenario.

The main contribution of this paper is to develop the hidden human target detection model inspired by physiological characteristics. The addition of physiological signal dimension improves the detection rate of hidden human targets. Previously, researchers acquire physiological signals after detecting human targets, and we innovatively use physiological signals as a class of features for human target detection. Specifically, to remove a large number of candidate boxes generated by classical selective search methods, the improved generation strategy constrained by scale prior and mutual correlation of physiological signal is proposed. Simultaneously, we aggregate the skin detection model to characterize the difference between the spatial pixel distribution of background human target region for suppressing the background candidate boxes. The performance of the raised model is verified by the homemade dataset.

2. Proposed model framework

Fig. 2 shows the pipeline of hidden human detection model inspired by physiological characteristics [23–25]. First, a number of proposals are generated for possible location. Second, the qualified candidate boxes are generated using scale prior constraint and mutual correlation of physiological signal. Third, the proposals located in the background are suppressed by aggregating the image features using skin detection model. Finally, the targeted location information is output to obtain the hidden human target location.

3. Proposal generation inspired by physiological signals

The traditional selective search approach only takes four aspects into account namely color, texture, image size, and regional fitness. Different from other non-biological objects to be detected, the human targets have unique characteristics in temporal dimensionality owing to physiological activities. Enlightened by this phenomenon, we propose the novel selective search method considering not only spatial features but also temporal features which is connected with physiological activities.

The classical selective search method was proposed by Sande et al. [26], and was widely used in two-stage target detection algorithms such as Fast RCNN. The definition of the classical selective search method is as follows [27].

$$s\left(r_i, r_j\right) = a_1 s_{colour}\left(r_i, r_j\right) + a_2 s_{texture}\left(r_i, r_j\right) + a_3 s_{size}\left(r_i, r_j\right) + a_4 s_{fill}\left(r_i, r_j\right) \tag{1}$$

where $s_{colour}\left(r_i, r_j\right)$, $s_{texture}\left(r_i, r_j\right)$, $s_{size}\left(r_i, r_j\right)$ and $s_{fill}\left(r_i, r_j\right)$ represent the color similarity, texture similarity, size similarity, and fill similarity between regions, respectively.

Based on above formula, the traditional selective search method only takes four aspects into account namely color, texture, size and filling degree. The physiological characteristics are not considered, and the size prior constraint for hidden human region is not introduced.

Physiological signal is the time sequence signal, and the features used by the classical selective search are scalar. Therefore, we transform physiological signal into scalar signal by Fourier transform, and then apply it for candidate box generation. In this task, the scalar signal is the heart rate value of the hidden human target calculated by IPPG.

In addition, the traditional selective search method produces thousands of candidate boxes at a time. If this paper calculates the physiological signals for these candidate boxes, the temporal and spatial complexity is tremendous. To reduce computational complexity, we bring in the scale priori and physiological characteristics into the traditional selective search method to suppress the candidate boxes. The definition of the proposal generation method enlightened by physiological signals is as follows.

$$s(r_i, r_j) = a_1 * s_c(r_i, r_j) + a_2 * s_t(r_i, r_j) + a_3 * s_s(r_i, r_j) + a_4 * s_f(r_i, r_j) + a_5 * s_r + a_6 * s_t \tag{2}$$

where $a_i \in \{0, 1\}$ indicates whether similarity is used or not. The proposed candidate box generation strategy considers six aspects namely color, texture, size, match, aspect ratio, and heart rate threshold. When $a_i = 1$, it means the corresponding similarity is considered. $s_c(r_i, r_j)$, $s_t(r_i, r_j)$, $s_s(r_i, r_j)$, $s_f(r_i, r_j)$, $s_r(r_i, r_j)$, and $s_t(r_i, r_j)$ represent color similarity, texture similarity, size similarity, fill similarity, aspect ratio information, and heart rate threshold, respectively. Based on the region similarity $s(r_i, r_j)$ defined above, the strategy combines two similar regions together and separates two different regions.

s_r is defined as the priori constraint for generating candidate boxes. If s_r of the candidate box is less than 2, it is taken into account, otherwise it is filtered out. To accurately obtain this prior value, we performed statistical analysis on ImageNet, which is the large image dataset, currently with 14,197,122 images [28]. This publicly available dataset has now become the fundamental project in the field of computer vision and deep learning research, and numerous studies have been conducted on this dataset [29–32]. Therefore, s_r obtained from statistical analysis based on ImageNet is considered to be representative.

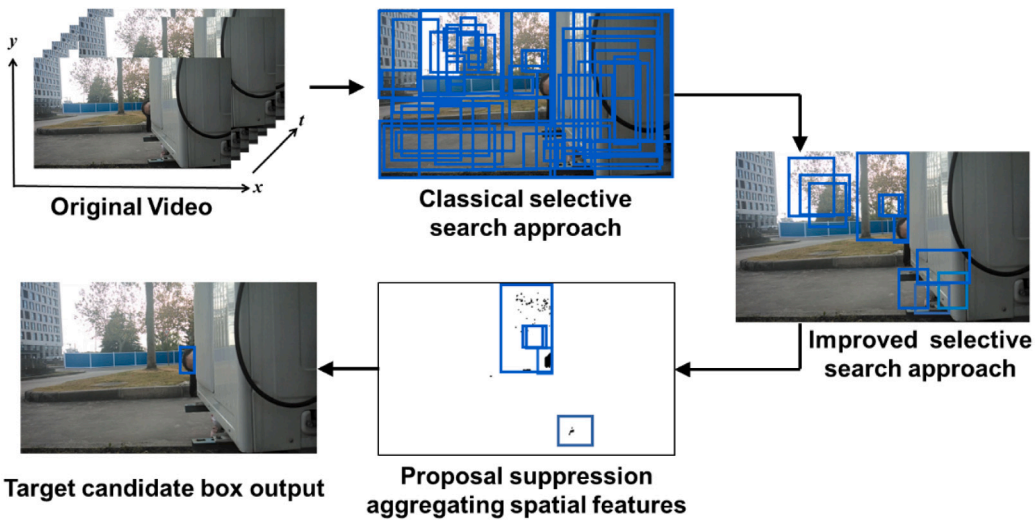

Fig. 2. The proposed model framework for hidden human target detection. A series of candidate boxes are first generated by classical selective search approach, then qualified by the prior knowledge constraints and mutual correlation analysis, followed by the suppression of the candidate boxes by aggregating the spatial features of the images with the skin detection model, and finally the target location information is output to obtain the hidden human target location.

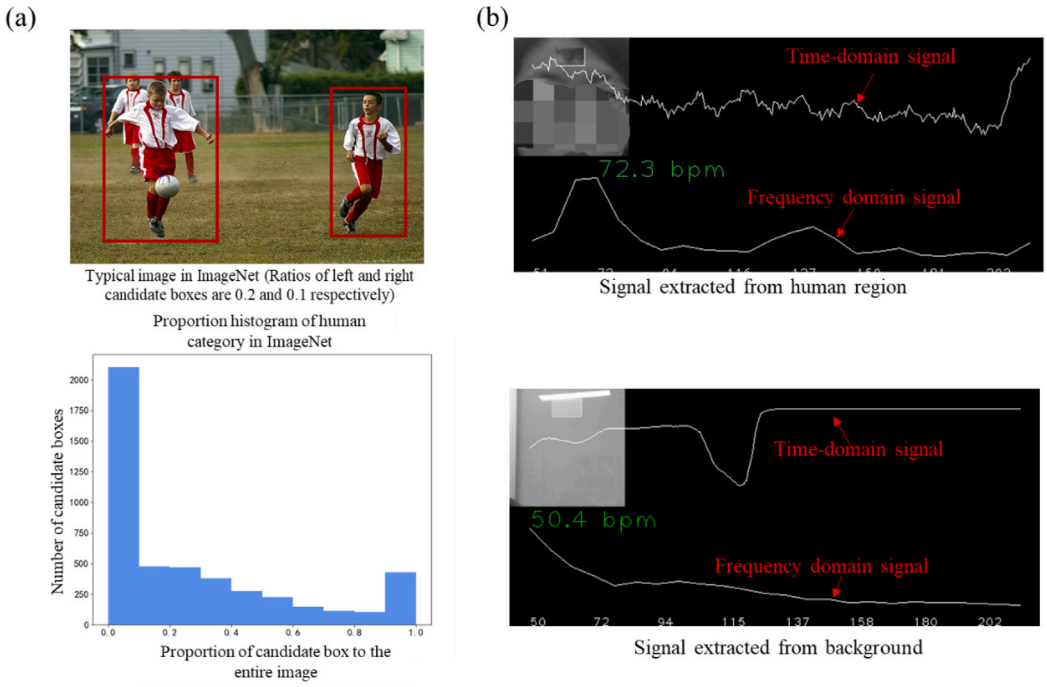

Fig. 3. Description of the values of s_r (a) and s_t (b). The left part of Fig. 3 (a) is the distribution of candidate boxes for human category from ImageNet: the upper half is the candidate boxes of the typical image for human body in ImageNet, where the ratio of human candidate boxes on the left and right are 0.2 and 0.1, respectively; the lower half is the histogram of the size proportion distribution of the candidate boxes from ImageNet, and it can be seen that the size of candidate boxes for human category is mostly concentrated in the ratio of 0.1, and this size is the smallest region of human target in ImageNet. The size of 0.1 is still much larger than the hidden human target, indicating that the size of 0.1 will not miss the hidden human target. Hence, 0.1 is assigned to s_r. The right part of Figure (b) shows the typical time and frequency domain plots of the signals of the human body and the background. The upper half is the heart rate signal variation in the forehead region of human body, and it mainly locates between 65–90 bpm; the lower half is the signal variation of the background, and it mainly locates around 50 bpm. The number of heart rate between 65 bmp and 90 bmp is counted for one video as number ratio s_t.

We collect information about the bounding boxes of human category in ImageNet, and compute the proportions of these candidate boxes to the entire image. The proportion histogram of human category is shown in Fig. 3(a). From the statistical figure, we can obtain the size range of candidate boxes, and further acquire the size limit for the candidate boxes to be generated in this task. As shown in Fig. 3(a), the size of candidate boxes for human category is mostly concentrated in the size of 0.1, and this size is the smallest region of human target in ImageNet. The size of 0.1 is still much larger than the hidden human target, indicating that the size of 0.1 will not miss the hidden human target. Hence, this prior value is very reasonable, and the candidate boxes whose size is larger than 0.1 are eliminated.

In terms of s_t, the current literature shows that the human heart rate is mainly in the range from 65 bpm to 90 bpm [33,34]. As shown in Fig. 3(b), there is the large difference in heart rate signal between human part and background. For one video, the number of heart rate between 65 bmp and 90 bmp is counted. If the number ratio in this

heart rate range exceeds 0.9, it is judged as human, otherwise it is judged as background.

Based on these six angles, a number of candidate boxes that may contain hidden body areas are generated.

4. Proposal suppression strategy based on physiological feature

IPPG is applied for extracting physiological signal. First, the original video is decomposed into R, G, and B three channels. The G channel is chosen to extract the time-varying signal. Then this paper performs Fourier transform to acquire the blood volume pulse waveform. The peak of the energy spectrum is found to gain the heart rate for the following analysis.

Before extracting the heart rate, the raw signals are normalized. The formula is as follows.

$$x(n) = x(n) - \frac{1}{N}\sum_{n=0}^{N-1} x(n) \tag{3}$$

where $x(n)$ represents the timing signal, and its signal length is N.

It is an assumption that the signal in the Hamming window is changing periodically. For this reason, the Hamming window of 15–20 s is applied for processing the signal to obtain the periodic signal. In the current work, we set Hamming window length to 250.

$$W(n,\alpha) = (1-\alpha) - \alpha * \cos\left(\frac{2 * \pi * n}{N-1}\right) \tag{4}$$

In this equation, α is usually set as 0.46.

After adding the Hamming window, the FFT transform is performed using the following equation to obtain the frequency domain signal.

$$X_k = \sum_{n=0}^{N-1} x_n e^{\frac{-2\pi jkn}{N}} \tag{5}$$

where x_n is the average value of the three channel pixels, which varies with time n, X_k is the result of the FFT transform in the frequency domain. The main problem of IPPG technology is the interference caused by motion artifacts. For example, when the pixel changes in candidate boxes are not influenced by the reflected light of the heart rate signal yet by the external motion, the quality of the heart rate signal will be substantially degraded. The Independent Component Analysis (ICA) is applied for reducing the signal quality deterioration caused by this external motion interference [35–37].

ICA presumes that the observed RGB three channels are linear mixture of original sources, i.e. $x_i(t) = \sum_{j=1}^{3} a_{ij} * s_j(t)$, in which $x_i(t)$ is the observed signal, $x_j(t)$ is the source signal whose distribution is non-Gaussian, and a_{ij} referring to different weights, can be re-expressed by matrix. The a_{ij} can be calculated according to ICA algorithm.

$$X(t) = AS(t) \tag{6}$$

The goal of this paper is to find the demixing matrix W to recover the source signal, which can be expressed as follows.

$$\hat{s}(t) = WX(t) \tag{7}$$

where $\hat{s}(t)$ the approximation of the inverse of source signal.

By using ICA, this paper can acquire the relatively pure signals and remove the interference of ambient noise in the same frequency range. The typical results are shown in Fig. 4. We can find that the shape of decomposed signals is closer to that of real heart beat signal, and the waveform is more periodic. This also demonstrates that ICA can reduce the interference of motion artifacts to a certain extent.

Besides, based on the shape of curves shown in Fig. 3(b), the spatio-temporal interdependent relationship of candidate boxes can be used for removing redundant bounding boxes. Correlation measures the similarity degree between two sequences. If the two signals are highly correlated, the corresponding two candidate boxes are considered as the same category. For instance, if two candidate boxes hit the human body but partially overlap, the signals extracted from these two candidate boxes should be highly correlated; if two candidate boxes respectively hit the background and human target regions, the correlation of these two signals is relatively small. The correlation between the signals is calculated as follows.

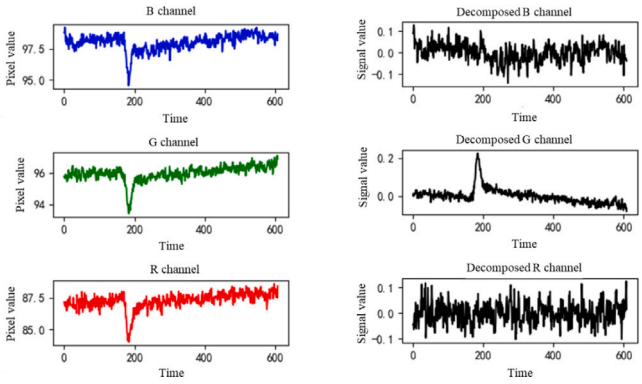

Fig. 4. Comparison of signals before and after applying ICA.

$$r_{xy}(l) = \sum_{-\infty}^{\infty} x(n)\, y(n-l) \tag{8}$$

where $x(n)$ and $y(n)$ are heart beat signals respectively extracted from two different candidate boxes.

The correlation coefficient matrix between the generated candidate boxes can be obtained. For hidden human target detection task, the background is the dominant region with large spatial occupation. By analyzing the correlation coefficient matrix, a small set of candidate boxes with larger correlations are selected for subsequent analysis, and the rest of candidate boxes are eliminated. To do this, the suppression of redundant candidate boxes can be effectively achieved.

5. Postprocessing aggregating spatial features

The classical selective search approach only considers the generic spatial features, and the inherent spatial features of human targets are not considered. Therefore, the skin detection algorithm [38] combined with image morphological manipulation is used to perform post-processing and thus improve the detection accuracy. The flowchart of post-processing and its corresponding processed images are shown in Fig. 5. The raw image is first binarized using threshold segmentation method in HSV and YCbCr color spaces, and these two binarized images are then subjected to the sum operation. Finally, the median filtering and open operation are respectively applied for reducing impulse pixels and bonded pixels. After skin detection, we check if the candidate box includes the area detected by the skin detection model. As long as any skin pixel falls in the candidate box, the candidate box is considered to be valid human body. If none of the skin pixels in the candidate box, we consider the candidate box as the background and filter it. The candidate box of background can be filtered via the postprocessing, and thus improving the detection accuracy of the hidden human target.

First, the raw RGB image is binarized in HSV and YCbCr color spaces. Second, the two binarized images are subjected to the sum operation. Finally, the median filtering and open operation are used to obtain the mask image for the subsequent analysis.

6. Results and discussion

6.1. Hidden human target dataset

To design possible human hiding scenes, the common objects such as walls, cars, trees, curtains, and tables are selected as the shelters. The volunteers, four men and four women, aged 20–40, crouched and

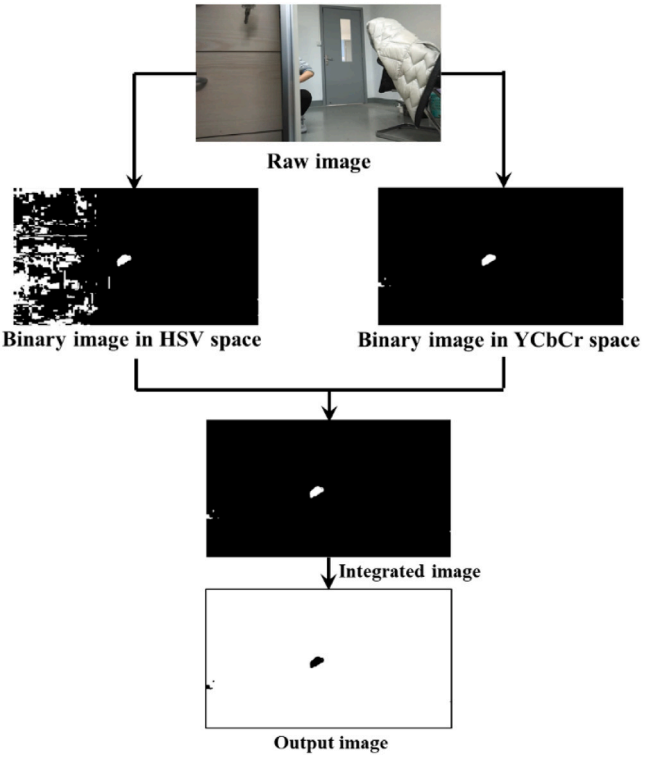

Fig. 5. Flowchart of post-processing and its corresponding processed images.

Table 1
Model performance comparison.

Model	Accuracy			IoU		
	Indoor	Outdoor	Overall	Indoor	Outdoor	Overall
Our model	64%	44%	54%	0.19	0.05	0.12
YOLO v4*	45%	5%	25%	0.17	0.08	0.11
YOLO v4	34%	28%	31%	0.15	0.09	0.11
Haar	36%	27%	32%	0.12	0.13	0.12
LBP	16%	9%	12%	0.05	0.07	0.06
HOG+SVM	30%	11%	21%	0.14	0.04	0.09

Note: YOLO v4* is fine-tuned on our homemade dataset.

stood motionless as they hid behind these places, with a small part of their bodies exposed. The RGB camera was used for imaging, with the shooting time of 20–30 s, the camera frame rate of 30 frames per second, and the image resolution of 1520 pixels × 720 pixels. Considering the influence of ambient light, the experiment was set up with the indoor scene and outdoor scene. A total of 100 videos were collected, including 50 indoor scenes and 50 outdoor scenes. Fig. 6 shows the typical images in homemade hidden human target dataset.

The experiment has been approved by the East China Normal University Committee on Human Research Protection Standard Operation Procedures (protocol code HR 087–2019).

6.2. Model performance

The mean Average Precision (mAP) is usually employed to measure model performance in object detection task. In this task, the confidence information cannot be determined. Therefore, the mAP metric cannot be derived. For hidden human target detection task, the accuracy and Intersection over Union (IoU) are used to verify the model performance. If the remaining candidate boxes hit the hidden body and the number of remaining candidate boxes is less than 5, we consider the detection as successful, otherwise we consider the detection as unsuccessful. The accuracy metric is obtained by counting the success and failure of the test data. IoU refers to the intersection ratio of groundtruth box and candidate box.

Table 1 shows the model performance comparison for hidden human target detection task. As shown in Table 1, the accuracies of the proposed model on indoor scene, outdoor scene, and overall dataset are 64%, 44%, and 54%, respectively, which are 19%, 39%, and 29% higher than the fine-tuned YOLO v4 model. With respect to IoU metric, the detection IoU values of the proposed model are 0.19, 0.05, and 0.12 for indoor scene, outdoor scene, and overall dataset, respectively. This is not much different from the fine-tuned YOLO v4 model. Therefore, the proposed model can improve detection accuracy while maintaining the relatively high IoU. Compared to the models based on traditional detection operators, the proposed model has better performance, with the overall accuracy (IoU) of 54% (0.12) versus 32% (0.12), 12% (0.06), and 21% (0.09) for Haar, LBP, HOG features, respectively. In addition, we used the video object detection methods such as LSTM-SSD [39] in this task, but the gradient explosion occurred during the training process. This may be because compared to the data in video object detection dataset, the videos in our homemade dataset do not have significant spatial displacement and have insignificant spatial features and rather weak temporal features. These features are not learned by existing video object detection algorithms.

The algorithm is currently still in the validation phase and is not yet capable of processing in real-time. For videos with a single frame resolution of 1520 pixels × 720 pixels and a duration of 20–30 s, the processing time is approximately between 30 to 60 s. In further study, we plan to reduce the processing time of the algorithm using image processing methods such reducing the spatio-temporal resolution of the videos.

6.3. Influence of application scenes on model performance

The detection accuracy of indoor scene is 20% higher than that of outdoor scene (Table 1), indicating that the proposed model is vulnerable to the influence of light. The influence factors related to the sunlight such as the uniform radiation conditions will bring the changes to the videos of outdoor scene, and they will cause the degradation of detection performance. As shown in Fig. 7, in the indoor scene, the signal related to the human body is less disturbed by high frequency noise; in the outdoor scene, the high frequency component of the signal increases. Therefore, the signal related to the human body is more seriously disturbed by light noise in the outdoor environment. In future research, the influence of illumination changes on model performance can be reduced from algorithmic or hardware perspective.

6.4. Influence of candidate box size on model performance

The size of candidate box affects the quality of the extracted signals. As shown in Fig. 8, when the size of candidate box increases, the high frequency component increases and in turn influences the quality of the extracted signal.

The similar phenomenon is also observed for outdoor scenes. As shown in Fig. 9, when the size of candidate box increases, the signal receives more interference. Unlike the indoor scene, the interference from the outdoor scene not only adds the noise to the original signal, but also changes the shape of the original signal in the time domain.

Whether it is the outdoor scene or the indoor scene, the increase of the percentage of background signal in the signal of human body is the main reason for the change of the original signal of human body. Compared to the indoor scene, the background region of outdoor scene contains more interference factors. This is one of the reasons why the raw signal of human body in outdoor scene changes a lot as the candidate box size increases. In further research, the extraction method of original signal needs to be improved to reduce the influence of the size of candidate boxes. The use of unmixing methods is also important to reduce this effect.

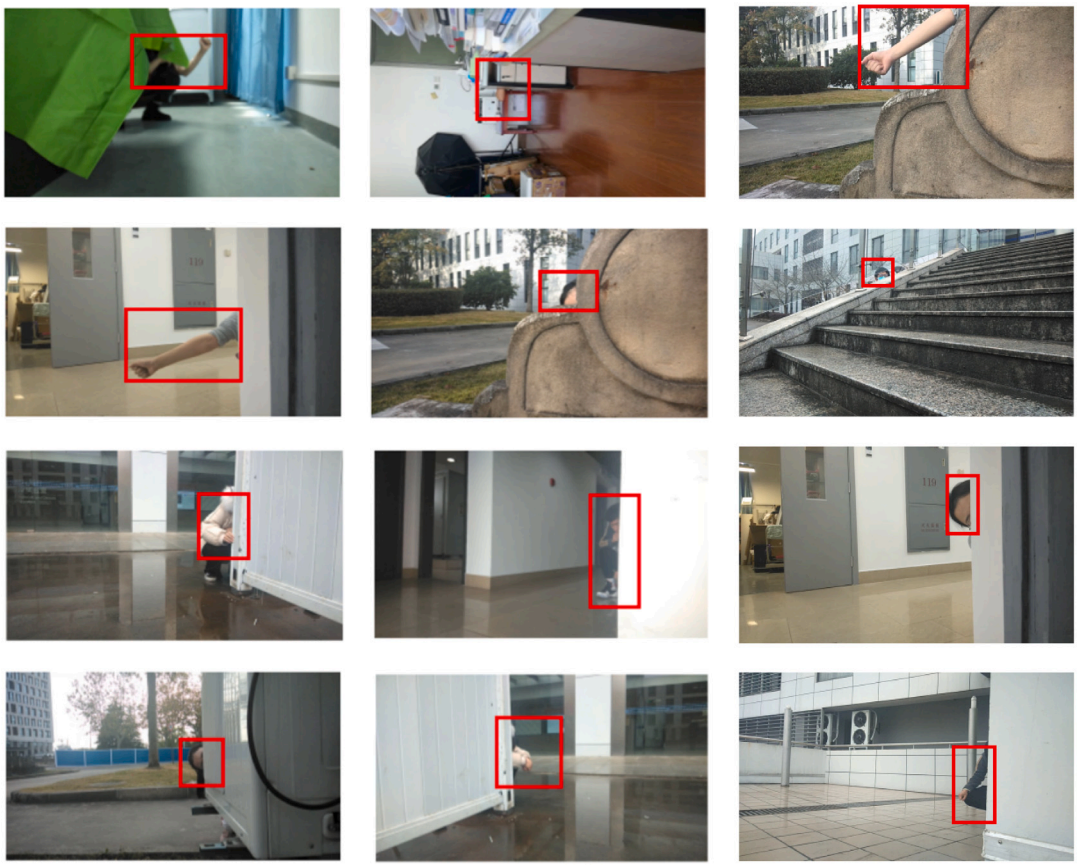

Fig. 6. Typical images in homemade hidden human target dataset. The area in red box is the location of hidden human target.

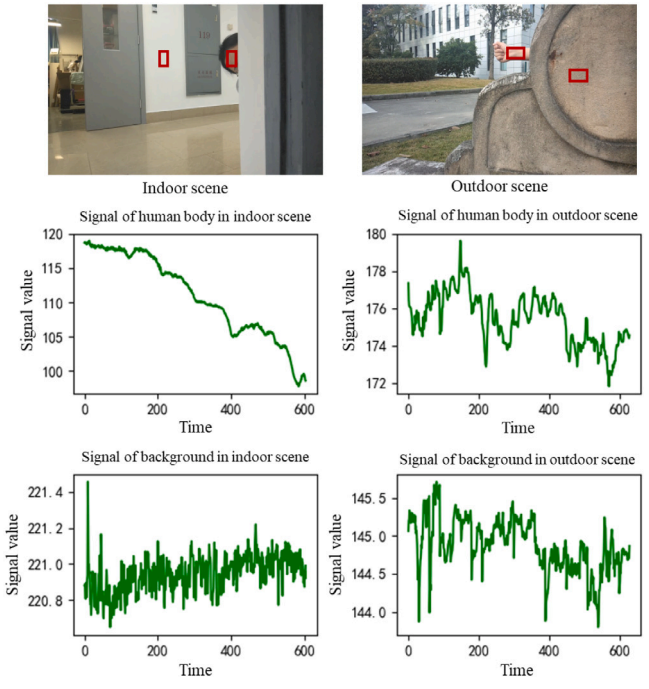

Fig. 7. Signals of human body and background in indoor scene (left column) and outdoor scene (right column). The red bounding boxes mean the selected regions for analysis.

6.5. Other influence factors

When the human body is far away from the camera, the detection performance of the proposed model decreases because the quality of the physiological signal inherent to the human target is severely degraded. The detection performance is also related to the exposed body parts. The extraction of physiological signal is difficult for human body regions with thick fat layer and not rich capillaries. The existing studies show that the forehead and wrist are the best candidates for detecting heart rate signals [40]. In addition, the presence of non-stationary objects such as shaking leaves in background will have the impact on the detection performance of the proposed model.

6.6. Ablation experiments

To probe the role of each module, this paper conducted the ablation experiments. Table 2 shows the results of ablation experiments. In Table 2, for outdoor scenes, the introduction of physiological feature contributes more efforts to improving model accuracy than other modules. For indoor scenes, the skin detection and physiological feature contribute more efforts to raising model performance. From IoU metric, both interdependent relationship and physiological feature are able to improve model performance. Overall, the addition of each module brings gains in model detection performance.

7. Conclusion

To address the performance degradation of general object detection models and traditional human target detectors in the hidden human target detection task, the physiological characteristics inspired hidden human object detection model was proposed in the current work. The experimental results on homemade dataset show that the detection accuracies of the proposed model are 64%, 44%, and 54% for indoor

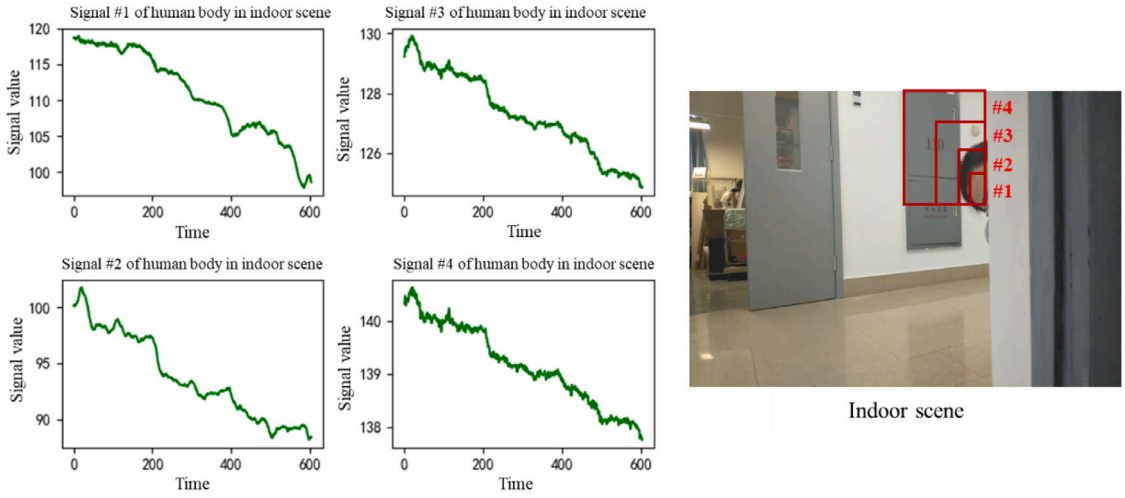

Fig. 8. Signals extracted from candidate boxes with four different sizes in indoor scene. The candidate boxes are labeled as #1, #2, #3, and #4 from small to large, respectively.

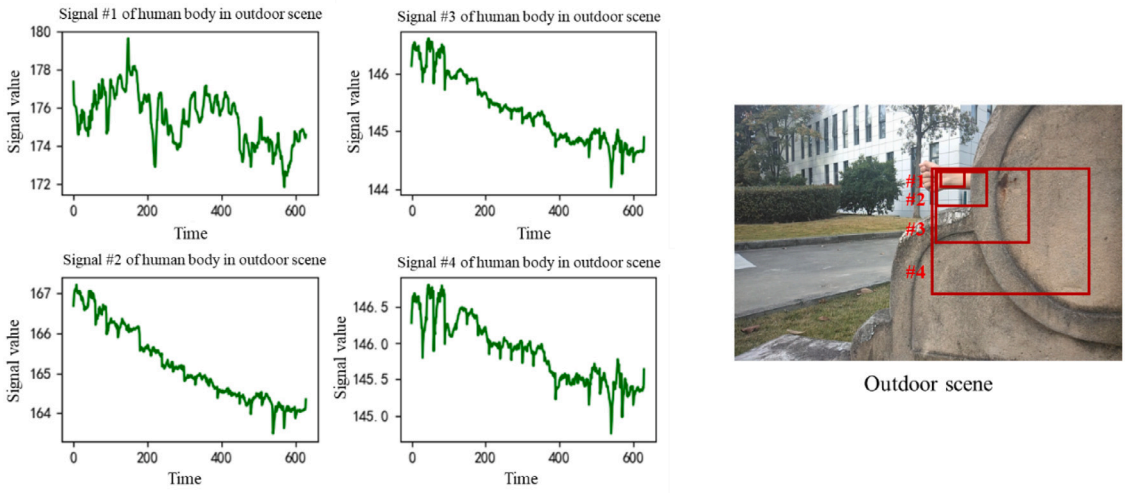

Fig. 9. Signals extracted from candidate boxes with four different sizes in outdoor scene. The candidate boxes are labeled as #1, #2, #3, and #4 from small to large, respectively.

Table 2
Results of ablation experiments for our model.

Interdependent relationship	Physiological feature	Skin detection	Accuracy			IoU		
			Indoor	Outdoor	Overall	Indoor	Outdoor	Overall
✓	✓	✓	64%	44%	54%	0.19	0.05	0.12
×	×	✓	36%	34%	35%	–	–	–
×	✓	✓	44%	40%	42%	0.11	0.04	0.07
✓	×	✓	54%	18%	36%	0.14	0.01	0.07

scene, outdoor scene, and overall dataset, respectively. The proposed model outperforms YOLO v4 models, and the models based on HOG, LBP, and Haar features, with at least 22% improvement in detection accuracy, demonstrating the feasibility of using the temporal–spatial physiological characteristics for the assistance in hidden human target detection. In the future, the model is expected to play an important role in military rescue, public security investigation and other fields. In addition, this method can be applied to detect animal targets or concealed animal targets. Like humans, animals (especially mammals) also exhibit physiological signals. Although some animals have fur-covered skin, we can still extract physiological signals through their partially exposed skin or through the respiratory-induced subtle body movement signals, which can then be used to detect animal targets or concealed animal targets.

CRediT authorship contribution statement

Menghan Hu: Conceptualization, Methodology, Supervision, Writing – review & editing. **Lejing Zhang:** Investigation, Data curation, Software, Validation, Writing – original draft. **Bailiang Zhao:** Writing – review & editing. **Yunlu Wang:** Supervision. **Qingli Li:** Supervision. **Lianghui Ding:** Supervision. **Yuan Cao:** Supervision.

Declaration of competing interest

The authors declare that there is no conflict of interests regarding the publication of this paper.

Data availability

Data will be made available on request.

Acknowledgments

This work is sponsored by the National Natural Science Foundation of China (No. 62371189).

References

[1] Z. Zakria, J. Deng, R. Kumar, M.S. Khokhar, J. Cai, J. Kumar, Multiscale and direction target detecting in remote sensing images via modified YOLO-v4, IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 15 (2022) 1039–1048.
[2] F. Yan, Y. Xu, Improved target detection algorithm based on YOLO, in: International Conference on Robotics, Control and Automation Engineering (RCAE), IEEE, 2021, pp. 21–25.
[3] Y. Qing, W. Liu, L. Feng, W. Gao, Improved yolo network for free-angle remote sensing target detection, Remote Sens. 13 (11) (2021) 2171.
[4] S. Song, Z. Miao, H. Yu, J. Fang, K. Zheng, C. Ma, S. Wang, Deep domain adaptation based multi-spectral salient object detection, IEEE Trans. Multimed. 24 (2022) 128–140.
[5] N. Huang, Y. Yang, D. Zhang, Q. Zhang, J. Han, Employing bilinear fusion and saliency prior information for RGB-D salient object detection, IEEE Trans. Multimed. 24 (2022) 1651–1664.
[6] L. Ruan, Y. Han, J. Sun, Q. Chen, J. Li, Facial expression recognition in facial occlusion scenarios: A path selection multi-network, Displays (2022) 102245.
[7] L. Zhang, B. Verma, D. Tjondronegoro, V. Chandran, Facial expression analysis under partial occlusion: A survey, ACM Comput. Surv. 51 (2) (2018) 1–49.
[8] J. Niu, X. Wang, D. Wang, L. Ran, A novel method of human joint prediction in an occlusion scene by using low-cost motion capture technique, Sensors 20 (4) (2020) 1119.
[9] M. Zhang, R. Zhang, Y. Yang, H. Bai, J. Zhang, J. Guo, Isnet: Shape matters for infrared small target detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 877–886.
[10] F. Bousefsaf, D. Djeldjli, Y. Ouzar, C. Maaoui, A. Pruski, iPPG 2 cPPG: Reconstructing contact from imaging photoplethysmographic signals using U-net architectures, Comput. Biol. Med. 138 (2021) 104860.
[11] R.H. Goudarzi, S.S. Mousavi, M. Charmi, Using imaging photoplethysmography (IPPG) signal for blood pressure estimation, in: International Conference on Machine Vision and Image Processing (MVIP), IEEE, 2020, pp. 1–6.
[12] L. Zhang, Y. Wang, M. Hu, Q. Li, Hidden human target detection model inspired by physiological signals, in: International Forum on Digital TV and Wireless Multimedia Communications, Springer, 2022, pp. 229–238.
[13] M. Hu, F. Qian, D. Guo, X. Wang, L. He, F. Ren, ETA-rppgnet: effective time-domain attention network for remote heart rate measurement, IEEE Trans. Instrum. Meas. 70 (2021) 1–12.
[14] A. Comas, T.K. Marks, H. Mansour, S. Lohit, Y. Ma, X. Liu, Turnip: Time-series U-net with recurrence for NIR imaging PPG, in: IEEE International Conference on Image Processing (ICIP), IEEE, 2021, pp. 309–313.
[15] A. Pai, A. Veeraraghavan, A. Sabharwal, HRVCam: robust camera-based measurement of heart rate variability, J. Biomed. Opt. 26 (2) (2021) 022707.
[16] D.G. Lowe, Object recognition from local scale-invariant features, in: Proceedings of the IEEE International Conference on Computer Vision, Vol. 2, IEEE, 1999, pp. 1150–1157.
[17] C.P. Papageorgiou, M. Oren, T. Poggio, A general framework for object detection, in: International Conference on Computer Vision, IEEE, 1998, pp. 555–562.
[18] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), Vol. 1, IEEE, 2005, pp. 886–893.
[19] Y. Song, Z. Xie, X. Wang, Y. Zou, MS-YOLO: Object detection based on YOLOv5 optimized fusion millimeter-wave radar and machine vision, IEEE Sens. J. (2022) 1.
[20] Y.-h. Li, X. Tan, W. Zhang, Q.-b. Jiao, Y.-x. Xu, H. Li, Y.-b. Zou, L. Yang, Y.-p. Fang, Research and application of several key techniques in hyperspectral image preprocessing, Front. Plant Sci. 12 (2021) 627865.
[21] R. Zhang, H. Li, K. Duan, S. You, K. Liu, F. Wang, Y. Hu, Automatic detection of earthquake-damaged buildings by integrating uav oblique photography and infrared thermal imaging, Remote Sens. 12 (16) (2020) 2621.
[22] C. Noviello, G. Esposito, I. Catapano, F. Soldovieri, Multilines imaging approach for mini-UAV radar imaging system, IEEE Geosci. Remote Sens. Lett. 19 (2021) 1–5.
[23] Y. Wang, M. Hu, Y. Zhou, Q. Li, N. Yao, G. Zhai, X.-P. Zhang, X. Yang, Unobtrusive and automatic classification of multiple People's abnormal respiratory patterns in real time using deep neural network and depth camera, IEEE Internet Things J. 7 (9) (2020) 8559–8571.
[24] A. Shokouhmand, S. Eckstrom, B. Gholami, N. Tavassolian, Camera-augmented non-contact vital sign monitoring in real time, IEEE Sens. J. 22 (12) (2022) 11965–11978.
[25] M.M. Shoushan, B.A. Reyes, A.R. Mejia Rodriguez, J.W. Chong, Contactless monitoring of heart rate variability during respiratory maneuvers, IEEE Sens. J. (2022) 1.
[26] K.E. Van de Sande, J.R. Uijlings, T. Gevers, A.W. Smeulders, Segmentation as selective search for object recognition, in: International Conference on Computer Vision, IEEE, 2011, pp. 1879–1886.
[27] J.R. Uijlings, K.E. Van De Sande, T. Gevers, A.W. Smeulders, Selective search for object recognition, Int. J. Comput. Vis. 104 (2) (2013) 154–171.
[28] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al., Imagenet large scale visual recognition challenge, Int. J. Comput. Vis. 115 (3) (2015) 211–252.
[29] S. Yun, S.J. Oh, B. Heo, D. Han, J. Choe, S. Chun, Re-labeling imagenet: from single to multi-labels, from global to localized labels, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 2340–2350.
[30] Y. Yousfi, J. Butora, E. Khvedchenya, J. Fridrich, ImageNet pre-trained CNNs for JPEG steganalysis, in: IEEE International Workshop on Information Forensics and Security (WIFS), IEEE, 2020, pp. 1–6.
[31] C. Xie, M. Tan, B. Gong, J. Wang, A.L. Yuille, Q.V. Le, Adversarial examples improve image recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 819–828.
[32] D. Li, H. Ling, S.W. Kim, K. Kreis, S. Fidler, A. Torralba, BigDatasetGAN: Synthesizing ImageNet with pixel-wise annotations, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 21330–21340.
[33] A. Ekström, K. Hellgren, A. Gräns, N. Pichaud, E. Sandblom, Dynamic changes in scope for heart rate and cardiac autonomic control during warm acclimation in rainbow trout, J. Exp. Biol. 219 (8) (2016) 1106–1109.
[34] C. Lefrançois, G. Claireaux, Influence of ambient oxygenation and temperature on metabolic scope and scope for heart rate in the common sole solea solea, Mar. Ecol. Prog. Ser. 259 (2003) 273–284.
[35] M.-Z. Poh, D.J. McDuff, R.W. Picard, Non-contact, automated cardiac pulse measurements using video imaging and blind source separation, Opt. Express 18 (10) (2010) 10762–10774.
[36] P. Ye, X. Wu, D. Gao, S. Deng, N. Xu, J. Chen, DP3 signal as a neuro-indictor for attentional processing of stereoscopic contents in varied depths within the 'comfort zone', Displays 63 (2020) 101953.
[37] Z. Gao, G. Zhai, H. Deng, X. Yang, Extended geometric models for stereoscopic 3D with vertical screen disparity, Displays 65 (2020) 101972.
[38] D. Dahmani, M. Cheref, S. Larabi, Zero-sum game theory model for segmenting skin regions, Image Vis. Comput. 99 (2020) 103925.
[39] H. Tang, Parabolic detection algorithm of tennis serve based on video image analysis technology, Secur. Commun. Netw. 2021 (2021).
[40] M.-Y. Chen, C.-W. Ting, A robust methodology for heartbeat detection in imaging photoplethysmography, in: International Conference on ICT Convergence (ICTC), IEEE, 2013, pp. 56–60.