IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received April 1, 2020, accepted April 20, 2020, date of publication April 27, 2020, date of current version May 21, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2990636

Using Deep Learning in Infrared Images to Enable Human Gesture Recognition for Autonomous Vehicles

KEKE GENG AND GUODONG YIN, (Senior Member, IEEE)
Mechanical Engineering Department, Southeast University, Nanjing 211189, China

Corresponding author: Guodong Yin (ygd@seu.edu.cn)

This work was supported in part by the National Natural Science Foundations of China under Grant 51905095 and Grant 51575103, in part by the Natural Science Foundations of Jiangsu under Grant BK20180401, and in part by the National Key Research and Development Program in China under Grant 2016YFB0100906 and Grant 2016YFD0700905.

ABSTRACT The realization of a novel human gesture recognition algorithm is essential to enable the effective collision avoidance of autonomous vehicles. Compared to visible spectrum cameras, the use of infrared imaging can enable more robust human gesture recognition in a complex environment. However, gesture recognition in infrared images has not been extensively investigated. In this work, we propose a model to detect human gestures, based on the improved YOLO-V3 network involving a saliency map as the second input channel to enhance the reuse of features and improve the network performance. Three DenseNet blocks are added before the residual components in the YOLO-V3 network to enhance the convolutional feature propagation. The saliency maps are obtained by multiscale superpixel segmentation, superpixel block clustering and cellular automata saliency detection. The obtained five scale saliency maps are fused using a Bayesian based fusion algorithm, and the final saliency image is generated. The infrared images composed of 4 main gesture classes are collected, each of which contains several approximated gestures in morphological terms. The training and testing datasets are generated, including original and augmented infrared images with a resolution of 640 × 480. The experimental results show that the proposed approach can enable real time human gesture detection for autonomous vehicles, with an average detection accuracy of 86.2%.

INDEX TERMS Human gesture recognition, autonomous vehicles, deep learning approach, infrared images, saliency maps.

I. INTRODUCTION

Human detection, as a key technology for autonomous vehicles, has attracted considerable research attention. In addition, human gesture recognition is necessary to predict the trend of human behavior, which is significantly important information to formulate effective collision avoidance strategies for autonomous vehicles.

At present, most methods for human gesture recognition involve the following processes: data collection, data preprocessing, feature quantity extraction and classifier learning, among which, feature extraction is the most important component. Researchers are constantly improving and developing this link to improve the accuracy of gesture recognition models. To enable the feature extraction of human gestures, several researchers have proposed various methods. Hemati and Mirzakuchaki [1] focused on recognizing human gestures by considering the appearance (Harris features) and motion information (oriented optical flow) by constructing spatiotemporal features. Luvizon *et al.* [2] presented a framework for human gesture recognition, taking into account the depth maps of skeleton sequences by using spatial and temporal local features from the subgroups of joints aggregated using a robust method based on the VLAD algorithm and a pool of clusters. Murtaza *et al.* [3] used the features pertaining to the histograms of oriented gradients (HOG) to describe the motion history image (MHI) low dimensional representation to enable silhouette based view independent human gesture recognition. Ghamdi *et al.* [4] reported on the use of a space-time extension of the scale invariant feature transform (SIFT), which was originally applied to 2 dimensional (2D) volumetric images, for human gesture application.

The associate editor coordinating the review of this manuscript and approving it for publication was Chao Yang.

Murray *et al.* [5] proposed a multiview human gesture recognition model that relies entirely on the active acoustic sonar data to infer human action. Palhang *et al.* [6] addressed the problem of categorizing human gestures by devising bag of words models based on the covariance matrices of spatiotemporal features, with the features obtained using the histograms of optical flow. Dawn and Shaikh *et al.* [7] presented a comprehensive review regarding STIP based methods for human gesture recognition and concluded that STIP based detectors could robustly detect the interest points from video in the spatiotemporal domain. Li *et al.* [8] proposed a framework combining the fast HOG3D features and self-organization feature map (SOM) network to enable gesture recognition from unconstrained videos, thereby bypassing the demanding preprocessing required for human detection, tracking or contour extraction. However, these methods usually involve handcrafted feature extraction, which requires researchers to have a deep understanding of the acquired data features and feature extraction algorithms. In addition, in general, the performance of handcrafted feature extraction approaches is not sufficiently stable owing to the complexity of environmental factors (weather changes, illumination changes, background changes, etc.), and thus, human gesture recognition under such conditions is challenging when using traditional methods.

In recent years, the use of convolutional neural networks [9], [10] in visual recognition has become increasingly popular, and their excellent performance in such tasks has been demonstrated. To enable the feature extraction of human gestures, Kim *et al.* [11] proposed a modified convolutional neural network (CNN) having a three dimensional receptive field, to generate a set of feature maps from the human gesture descriptors derived from a spatiotemporal volume. Le *et al.* [12] presented a framework for human gesture recognition by using the temporal and spatial features extracted simultaneously by utilizing a fine to coarse (F2C) CNN architecture optimized for human skeleton sequences. Wang *et al.* [13] proposed a visual attribute augmented 3D CNN framework that integrated the visual attributes (including detection, encoding and classification) to enable gesture recognition in trimmed videos. Meng *et al.* [14] presented a hierarchical dropped CNN architecture with a dropped CNN (d-CNN) to extract deep human gesture features from a probabilistic speed insensitive color image; furthermore, the authors extended the d-CNN to a hierarchical structure (h-CNN), in which multiple scales of temporal information are encoded, to enhance the temporal discriminative power. Meng *et al.* [15] proposed a deep learning network for gesture recognition, which integrated a quaternion spatiotemporal convolutional neural network (QST-CNN) and long short term memory network (LSTM); in this approach, a quaternion expression for an RGB image was employed, and the values of the red, green, and blue channels were considered simultaneously as a whole in a spatial convolutional layer, thereby avoiding the loss of spatial features. Yang *et al.* [16] proposed a sequential convolutional neural network to extract the effective spatiotemporal features of human gesture from videos, thereby incorporating the strengths of both convolutional and recurrent operations. Li *et al.* [17] proposed an end to end deep convolutional neural network in which the skeleton sequences were transformed into images, and the spatial temporal information was learned to enable 3D human gesture recognition. Ji *et al.* [18] developed a novel 3D CNN model to enable gesture recognition by extracting features from both the spatial and temporal dimensions by performing 3D convolutions, thereby capturing the motion information encoded in multiple adjacent frames. Sun *et al.* [19] proposed a human gesture recognition approach by using factorized spatiotemporal convolutional networks that factorize the original 3D convolution kernel learning as a sequential process of learning the 2D spatial kernels in the spatial convolutional layers, followed by the learning of the 1D temporal kernels in the temporal convolutional layers. Bhattacharjee and Das [20] reported upon the exploration of a two stream convolutional neural network (2S-CNN) architecture involving the fusion of the dense optical flow features of the RGB frames and the salient object regions detected using a fast space-time saliency method to categorize the human gestures in videos.

In general, most of these methods are based on RGB images. However, visible light cameras require proper illumination to function effectively. In contrast, infrared cameras produce images based on the heat radiated by the human body; consequently, these cameras can be used regardless of the external illumination conditions, and they can overcome the influence of illumination changes, while still achieving satisfactory results under partial occlusion and overlap conditions. Figure 1 shows a comparison of the RGB images and infrared images with different human gestures in some specific scenes.

Figure 1 indicates that under the conditions of rain and foggy weather, the colors of the background and human clothes being similar, occlusion and overlap, evening, and night time, the features of the humans (contour, brightness, orientation, etc.) in infrared images have a higher degree of differentiation with respect to the background, compared to in RGB images. However, research regarding human motion recognition methods based on infrared images is still insufficient. Akula *et al.* [21] demonstrated the use of IR cameras in the field of ambient assisted living and discussed its performance in human gesture recognition. Li *et al.* [22] proposed a human gesture silhouette energy histogram algorithm by using the statistical background model and background subtraction method to extract the human gesture silhouettes to address the problem of night time human gesture recognition. Osada *et al.* [23] proposed a human gesture pattern monitor (HPM) constructed using a film infrared sensor (MFI), without employing a monitor camera to ensure the clients' privacy, as a tableware system.

The contribution of this work is threefold:

- Considering the low resolution of convolutional feature maps, three DenseNet blocks were added before the residual components in the YOLO-V3 network to

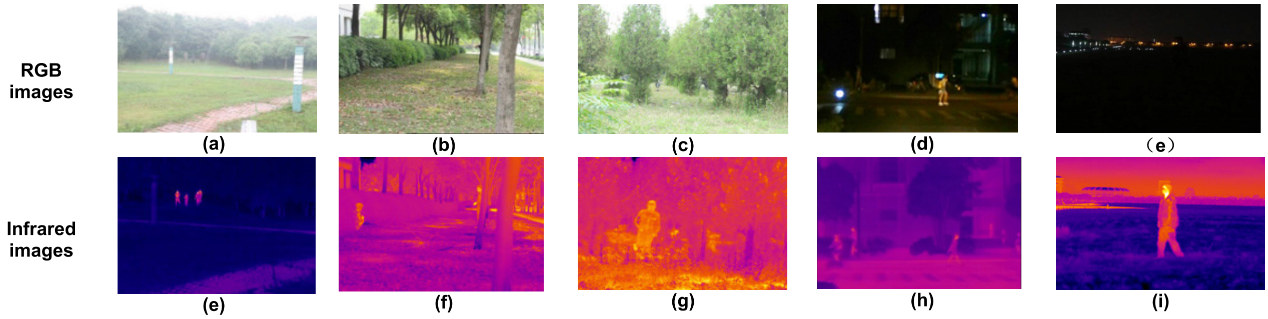


FIGURE 1. Comparison of human gestures in RGB images and infrared images under conditions of: (a) foggy weather; (b) occlusion; (c) background and targets having similar colors; (d) evening time with street light; (e) night time without street lights; (e)–(i) show the corresponding infrared images of the RGB images shown in (a)–(e).

enhance the convolutional feature propagation, thereby developing a novel human gesture detection network architecture;

- Infrared images lack information regarding the sharp edges and boundaries of variable human gestures. In addition, the temperature changes considerably influence the imaging results. To address this problem, the saliency maps of infrared images were detected as an additional input channel to the gesture detection network to improve the robustness of the proposed model;
- An infrared image dataset of human gestures in four main categories (‘‘squatting’’, ‘‘lying’’, ‘‘standing’’, and ‘‘walking’’) in an outdoor environment was generated for training and testing the human gesture detection network.

The remaining article is organized as follows. Section 2 describes the proposed human gesture detection algorithm, along with the improved YOLO-V3 algorithm and infrared image saliency detection algorithm. Section 3 describes the creation of the human gesture infrared image dataset, including the process of obtaining the original infrared image and image dataset enhancement methods. The experimental content and results are described in detail in Section 4. Section 5 summarizes the employed methods and conclusions of this work.

II. METHODOLOGY

The architecture of the proposed human gesture detection model for infrared images is illustrated in Figure 2.

The input of this human gesture detection model includes infrared images with a resolution of 640 × 480, as obtained using an infrared camera in outdoor environments. These images are processed and resized into grayscale images and saliency images, each with a resolution of 416 × 416, as the inputs of the proposed network. The feature maps from the two modalities are concatenated, and the 1 × 1 convolution operation is performed on the concatenated feature maps to reduce the dimensions and linearly merge the features. Furthermore, the parameters of the 1 × 1 convolution kernel are trained to reduce the dimensions of the concatenated feature maps to 52 × 52 × 128.

Considering the lower resolution of the convolutional layers in traditional YOLO-V3, three DenseNet blocks are added before the residual networks to improve the network performance from the perspective of feature reuse. The transfer function $H_i\,(i = 1, 2, 3, 4)$ employs the following network architecture: BN-ReLU-Conv, where BN denotes the batch normalization. The transfer functions H_1, H_2, H_3 and H_4 enable the nonlinear transformation of $x_0, [x_0, x_1], [x_0, x_1, x_2]$ and $[x_0, x_1, x_2, x_3]$ layers, respectively. The feature layer $[x_0, x_1, x_2, x_3, x_4]$ continues to propagate forward as the input of a transition layer. Finally, the feature layers are spliced into feature maps with resolutions of 52 × 52 × 256, 26 × 26 × 512, and 13 × 13 × 1024, and these feature maps are propagated forward.

The specific structure and parameters of the proposed human gesture detection network are shown in Figure 3.

A. HUMAN GESTURE DETECTION MODEL BASED ON THE IMPROVED YOLO-V3 NETWORK

In contrast to the faster R-CNN, which is a state of the art target detection and recognition network, the YOLO network generates both the coordinates and recurrence of each category directly through regression, which makes the YOLO network considerably faster than the faster R-CNN network. In the YOLO network series, the YOLO-V3 [26] network exhibits the highest detection accuracy, compared with those of the target detection network YOLO, YOLO-V2 [27] and SSD [28] networks.

The YOLO network simply divides the input images into an S × S grid. Each grid predicts the conditional probability C and bounding box B, each of which corresponds to five predicted values, including the center coordinates of bounding box (x and y), size of the image (*height* and *width*) and confidence score. The confidence can be obtained as

$$Conf = p_r \times IoU^t_{p.}, \quad p_r \in \{0,\ 1\}$$

where $p_r = 1$ if the object is in the grid, and $p_r = 0$ otherwise; IoU^t_p is used to denote the accuracy of the predicted bounding box relative to the ground truth. If the same target is detected by multiple bounding boxes, the bounding box with the highest score is selected by using nonmaximum suppression.

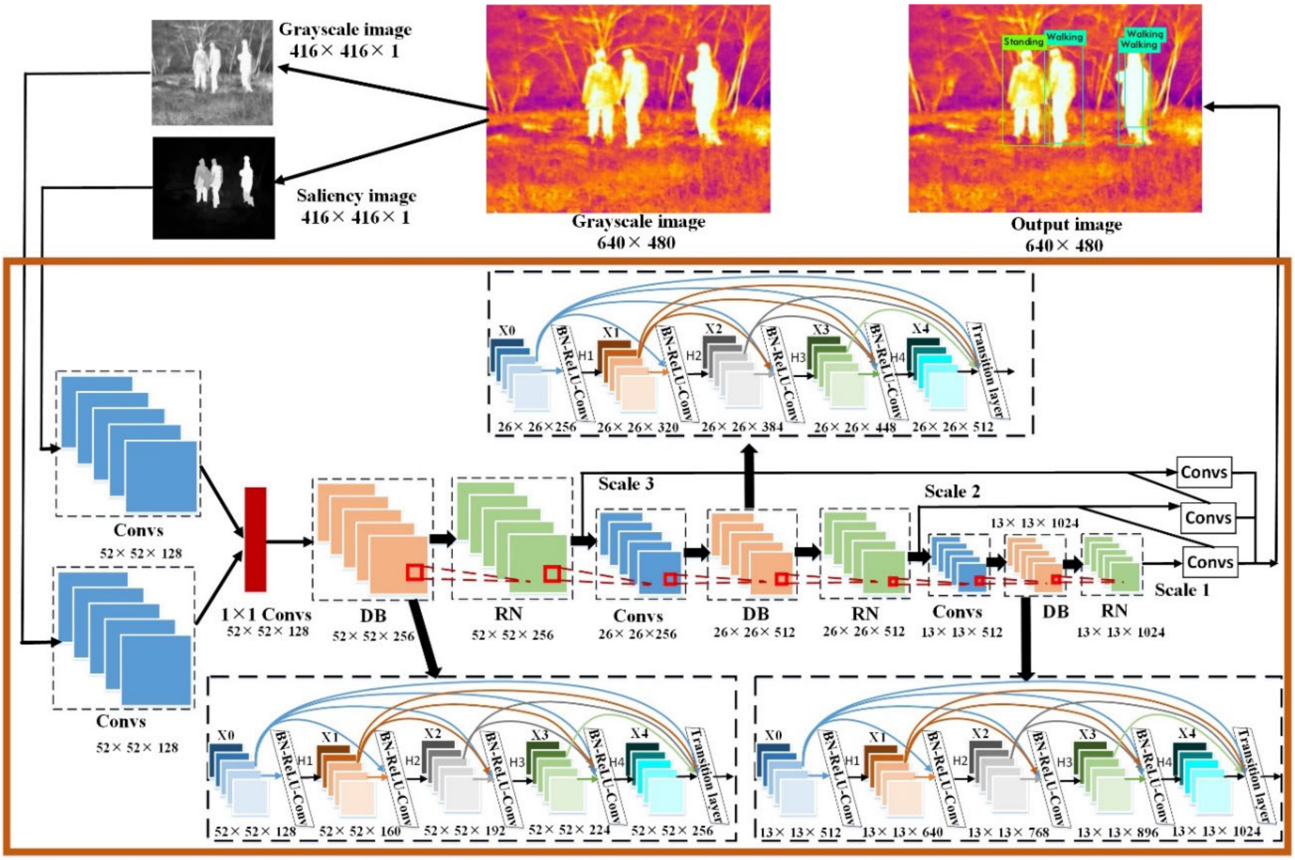


FIGURE 2. Architecture of proposed neural network: DB - DenseNet Block; Convs - Convolutional layers; RN - Residual network.

Although the YOLO network has a significant advantage in terms of the computational speed compared to the faster R-CNN, the accuracy of predicting the bounding box and classification is lower. To improve the object positioning accuracy and recall rate, the YOLO-V2 network implements an anchor box as in the faster R-CNN to improve the design of the network structure. Compared with the YOLO-V2 network, the most notable aspect regarding the YOLO-V3 is that a multiscale prediction method is employed, which leads to a qualitative improvement in the detection accuracy and average detection time.

Compared with the target objects in visible light images, the target objects in infrared images possess a considerably smaller amount of feature information. In the process of convolution and pooling of the YOLO-V3 network, a large amount of feature information is lost, which is unfavorable for the accurate localization and classification of targets. To address this problem, DenseNet [29] blocks are added before the residual components of Darknet-53, which is the basic network for the YOLO-V3. The use of the DenseNet improves the network performance in the context of feature reuse, which makes the feature information of the targets in infrared images more effective. The DenseNet blocks splice all the convolutional modules, owing to which, the input to each layer of the network includes the output of all the previous layers of the network. The use of such DenseNet blocks improves the transmission efficiency of the information and gradients in the network, as the gradient is obtained from the loss function and input signal. This network structure also enables the realization of regularization.

B. DETECTION OF SALIENCY MAP OF INFRARED IMAGES

Because infrared images lack information regarding the colors, textures, sharp edges and boundaries of variable human gestures, it is difficult to accurately distinguish the human body and background considering only the features of the brightness information in infrared images. In addition, changes in the temperature considerably influence the imaging results. To address this problem, the saliency maps of infrared images are detected as an additional input channel for the gesture detection network to improve the robustness of the proposed model. In this work, we propose an algorithm for the multiscale optimization of cellular automata to enable the detection of the saliency maps in infrared images, as shown in Figure 4.

First, the original infrared images are segmented into superpixel maps with five different scales by using the superpixel segmentation algorithm named simple linear iterative clustering (SLIC). Next, the numbers of superpixel blocks

	Type	Filters	Size	Input	Output
	Convolutional	32	3×3/1	416×416×1 (BC) 416×416×1 (SC)	416×416×32 (BC) 416×416×32 (SC)
	Convolutional	64	3×3/2	416×416×32 (BC) 416×416×32 (SC)	208×208×64 (BC) 208×208×64 (SC)
1	Convolutional	32	1×1/1	208×208×64 (BC) 208×208×64 (SC)	208×208×32 (BC) 208×208×32 (SC)
	Convolutional	64	3×3/1	208×208×32 (BC) 208×208×32 (SC)	208×208×64 (BC) 208×208×64 (SC)
	Residual			208×208×64 (BC) 208×208×64 (SC)	208×208×64 (BC) 208×208×64 (SC)
	Convolutional	128	3×3/2	208×208×64 (BC) 208×208×64 (SC)	104×104×128 (BC) 104×104×128 (SC)
2	Convolutional	64	1×1/1	104×104×128 (BC) 104×104×128 (SC)	104×104×64 (BC) 104×104×64 (SC)
	Convolutional	128	3×3/1	104×104×64 (BC) 104×104×64 (SC)	104×104×128 (BC) 104×104×128(SC)
	Residual			104×104×128 (BC) 104×104×128(SC)	104×104×128 (BC) 104×104×128 (SC)
	Convolutional	256	3×3/2	104×104×128 (BC) 104×104×128(SC)	52×52×256 (BC) 52×52×256 (SC)
	Convolutional	128	1×1/1	52×52×256 (BC) 52×52×256 (SC)	52×52×128
4	DenseNet DenseNet	32 32	1×1/1 3×3/1	52×52×128	52×52×(128+32)
	Output			52×52×128	52×52×256
8	Convolutional	128	1×1/1	52×52×256	52×52×128
	Convolutional	256	3×3/1	52×52×128	52×52×256
	Residual			52×52×256	52×52×256
	Convolutional	256	3×3/2	52×52×256	26×26×256
4	DenseNet DenseNet	64 64	1×1/1 3×3/1	26×26×256	26×26×(256+64)
	Output			26×26×256	26×26×512
8	Convolutional	256	1×1/1	26×26×512	26×26×256
	Convolutional	512	3×3/1	26×26×256	52×52×512
	Residual			26×26×512	26×26×512
	Convolutional	512	3×3/2	26×26×512	13×13×512
4	DenseNet DenseNet	128 128	1×1/1 3×3/1	13×13×512	13×13×(512+128)
	Output			13×13×512	13×13×1024
8	Convolutional	512	1×1/1	13×13×1024	13×13×512
	Convolutional	1024	3×3/1	13×13×512	13×13×1024
	Residual			13×13×1024	13×13×1024
	Avgpool		Global		
	Connected		1000		
	Softmax				

FIGURE 3. Network parameters of the improved YOLO-V3 network.

are reduced using the density based spatial clustering of application with noise (DBSCAN) algorithm. The superpixel maps of the saliency features are detected via the improved cellular automaton. Finally, using the framework of the fusion algorithm in Bayesian theory, the final saliency image is obtained.

1) SUPERPIXEL SEGMENTATION

The SLIC algorithm, which is an image segmentation algorithm, was proposed by Ren X. in 2003 [30]. This algorithm uses the similarity of the features between the pixels to group the pixels, and a small number of superpixels are used to describe the characteristics of the images. Consequently, the complexity and redundant information of images can be considerably reduced, which helps improve the speed of the subsequent image processing calculations and real time performance of target detection. The SLIC algorithm used in this paper is a simple linear iterative clustering algorithm, which transforms the original infrared image into a 5 dimensional feature vector.

The process employed by the SLIC superpixel segmentation algorithm can be described as follows: initialize the clustering center of the infrared images and set the number of superpixels and step size; calculate the gradient value of all the pixel points in the 3 × 3 field around the clustering center and reselect the clustering center according to the minimum gradient value; label each pixel in the neighborhood around

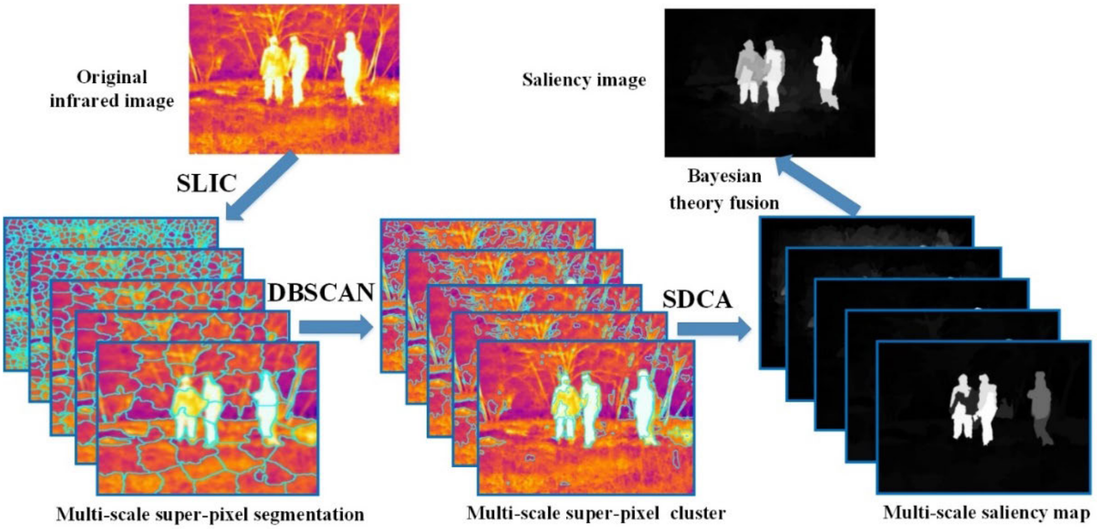


FIGURE 4. Frame diagram of the proposed saliency detection algorithm in infrared images.

each clustering center; calculate the distance between the pixel points in the surrounding neighborhood and its clustering center. The distance between two pixel points in the infrared image is obtained as

$$D' = \sqrt{\left(\frac{d_c}{N_c}\right)^2 + \left(\frac{d_s}{N_s}\right)^2}$$

where N_s represents the maximum spatial distance within the cluster; N_c represents the maximum color distance; d_c represents the color distance; d_s and represents the spatial distance.

The following scales were selected for the superpixel maps in this work: 50, 100, 200, 500, and 1000, as shown in Figure 4. Replacing the pixel information with the information of the superpixel blocks can effectively reduce the redundant information of the infrared images and increase the speed of the subsequent processing algorithms. Furthermore, by using different scales, human targets with different spatial scales in the infrared images can be segmented more effectively.

2) SUPERPIXEL CLUSTERING

DBSCAN, which is a clustering algorithm proposed by Ester M. [31], has become one of the most widely used clustering algorithms as it can discover arbitrarily shaped clusters and eliminate noisy data. The DBSCAN technique is a spatial data clustering method based on density, using which, a high density region can be divided into a cluster, and arbitrary shapes in the spatial dataset can be found. We present the following definitions to elucidate upon the mechanism of the algorithm:

- *Eps neighborhood of point p* : a circular area with the center at p and radius of Eps; p belongs to dataset D. The set of points included in the Eps neighborhood is denoted as $N_{Eps}(p) = \{q \in D \,|dist(p, q) \leq Eps\}$;
- *MinPts of point p* : minimum number of points in an *Eps neighborhood* of point *p*;
- *Density of point p*: number of points within the *Eps neighborhood* of point *p*;
- *Core point:* the point that satisfies the condition $|N_{Eps}(p)| \geq MinPts$; here, *MinPts* is the minimum number of points in the *Eps neighborhood* of point *p*;
- *Border point:* a point that is contained in the *Eps neighborhood* of point *p* but is not the core point;
- *Noise point:* a point that is neither a core point nor a border point;
- *Directly density reachable:* a point *p* is directly density reachable from a point *q* with regard to *Eps* and *MinPts*, if $p \in N_{Eps}(q)$ and $|N_{Eps}(q)| \geq MinPts$;
- *Density reachable:* a point *p* is density reachable from a point *q* with regard to *Eps* and *MinPts*, if there exists a chain of points $p_1, p_2, \cdots, p_n, p_1 = q, p_n = p$, such that p_{i+1} is directly density reachable from p_i;
- *Density connected:* a point *p* is density connected to a point q with regard to *Eps* and MinPts, if there exists a point w such that both p and q are density reachable from w.

The key concept of this algorithm is to find all the core points and form the clusters by clustering the core points with all the points that are reachable from it. The specific algorithm process can be described as follows: arbitrarily select a point *p* from the database; retrieve all the points *density reachable* from *p* with regard to *Eps* and *MinPts*; if *p* is a core point, a cluster is formed; if *p* is a border point, no points are *density reachable* from *p*, and DBSCAN visits the next point of the database; the process is continued until all the points have been processed.

3) SALIENCY DETECTION BASED ON CELLULAR AUTOMATA

In this paper, the cellular automaton is used to detect the saliency features of the infrared images. The clustered superpixel maps are taken as the input, and the accuracy of the saliency maps are improved by optimizing the update rules.

- *Impact Factor Matrix:* It is intuitive to accept that neighbors with more similar color features have a greater influence on a cell's next state. The similarity of any pair of superpixels is measured using a defined distance in the CIELAB color space. We construct the impact factor matrix $F = [f_{ij}]_{N\times N}$ by defining the impact factor f_{ij} of superpixel i to j as

$$f_{ij} = \begin{cases} \exp\left(\frac{-\left\|c_i, c_j\right\|}{\sigma_3^2}\right), & j \in NB\,(i) \\ 0\ , & i = j \textit{ or otherwise} \end{cases}$$

  where $\left\|c_i, c_j\right\|$ denotes the Euclidean distance in the CIELAB color space between the superpixel i and j; σ_3 is a parameter to control the degree of the similarity; $NB\,(i)$ is the set of neighbors of cell i. To normalize impact factor matrix, a degree matrix $D = diag\,\{d_1, d_2, \cdots, d_N\}$ is generated, where $d_i = \sum f_{ij}$. Finally, a row normalized impact factor matrix can be clearly determined as follows

$$F^* = D^{-1} \cdot F$$

- *Coherence Matrix:* Because the subsequent state of each cell is determined by its current state as well as the state of its neighbors, the importance of the two decisive factors must be balanced. In particular, if a superpixel is considerably different from all the neighbors in the color space, its next state will be primarily dependent on the cell itself. However, if a cell is similar to the neighbors, it is more likely to be assimilated by the local environment. We build a coherence matrix $C = diag\,\{c_1, c_2, \cdots, c_N\}$ to better promote the evolution of all the cells. The coherence of each cell toward its current state can be calculated as

$$c_i = \frac{1}{\max\left(f_{ij}\right)}$$

  To control $c_i \in [b, a+b]$, we construct the coherence matrix $C^* = diag\left\{c_1^*, c_2^*, \cdots, c_N^*\right\}$ using the following formulation:

$$c_i^* = a \cdot \frac{c_i - \min\left(c_j\right)}{\max\left(c_j\right) - \min\left(c_j\right)} + b$$

  where $j = 1, 2, \cdots, N$. We set the constants a and b as 0.6 and 0.2, respectively. Using the coherence matrix C^*, each cell can automatically evolve into a more accurate and steady state. Furthermore, the salient object can be more easily detected under the influence of the neighbors.
- *Synchronous Updating Rule:* In single layer cellular automata, all the cells update their states simultaneously according to the update rule. Given an impact factor matrix and coherence matrix, the synchronous updating rule $f : S^{NB} \to S$ can be defined as follows:

$$S^{t+1} = C^* \cdot S^t + \left(I - C^*\right) \cdot F^* \cdot S^t$$

  where I is the identity matrix, and C^* and F^* denote the coherence matrix and impact factor matrix, respectively. By using this update machine to create the original saliency map for each scale space, the respective optimized saliency maps can be obtained.

4) BAYESIAN THEORY FUSION METHOD

Due to the different scales of superpixel segmentation, the optimized saliency maps obtained in each scale space have their own advantages and disadvantages. This paper uses a fusion method based on the Bayesian theory, which can be used to obtain the optimal significance by combining the saliency values of each scale. In particular, the saliency map of any scale $S_i\,(i = 1, 2, \cdots, 5)$ is selected as the Bayesian prior probability, and the other four saliency maps $S_j\,(j \neq i, j = 1, 2, \cdots, 5)$ are defined as the likelihood probability. Let the current S_i merge with S_j separately. The final four posterior probability maps are added, and the average is considered as the final saliency map. The detailed steps can be described as follows:

- Use F_i and B_i to represent the foreground and background regions, respectively. N_{F_i} and N_{B_i} represent the number of pixels in the foreground and background regions, respectively.
- Calculate the distribution characteristics of S_j in the foreground and background regions. In the normalized statistical distribution histogram for the significance value S_j, the observation likelihood probability of pixel x can be expressed considering the value of the corresponding bit of $S_j\,(x)$, as follows:

$$\begin{cases} P\left(S_j\,(x) \middle| F_i\right) = \dfrac{N_{bF_i(S_j(x))}}{N_{F_i}} \\ P\left(S_j\,(x) \middle| B_i\right) = \dfrac{N_{bB_i(S_j(x))}}{N_{B_i}} \end{cases}$$

  where $N_{bF_i(S_j(x))}$ and $N_{bB_i(S_j(x))}$ respectively represent the number of pixels in the feature $S_j\,(x)$ bits falling in the foreground and background statistical histograms.
- If S_i is the Bayesian prior probability, the posterior probability can be calculated as follows:
- Add the results of the four time two-two fusion and average these values to obtain the final saliency map.

III. DATASET DESCRIPTION

A. IMAGE DATA ACQUISITION

The infrared images for training and testing the proposed deep learning human gesture recognition model were obtained using an infrared camera with a pixel resolution of 640 × 480. The images were acquired at the Southeast University Jiulong Lake Campus in Nanjing, Jiangsu, China. Our data acquisition platform consisted of an infrared camera and two RGB cameras, one of which had a polarizing filter, as shown in Figure 5. To facilitate the data collection process, the platform was fixed on top of an autonomous ground vehicle. The image registration of multiple cameras was achieved by using a calibration matrix.

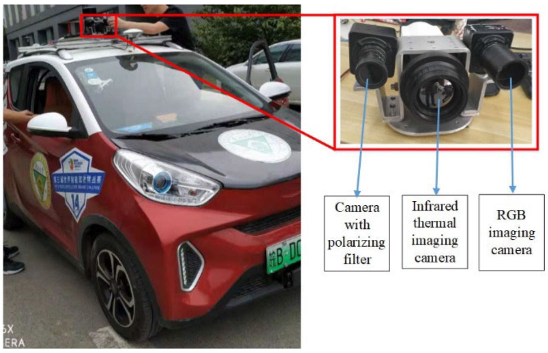

Camera with polarizing filter
Infrared thermal imaging camera
RGB imaging camera

FIGURE 5. Image data acquisition platform.

The infrared images were captured under sunny, cloudy, misty, and slightly rainy weather conditions (the light illuminance ranged from less than 50 lux to more than 50,000 lux). The acquisition period was relatively long and lasted from June to October. The acquisition was initiated at 9 a.m., 4 p.m., and 8 p.m. Furthermore, images were gathered in foggy weather and night time conditions.

A total of 2492 infrared images were initially collected, involving 4 main gesture classes, each of which contained several approximated gestures in morphological terms, including "squatting" (sitting, kneeling, squatting, etc.), "lying" (lying on the back, side, stomach, etc.), "standing" (facing the lens, with side to the lens, with back to the lens, etc.), "walking" (running, brisk walking, slow walking, etc.). Among all these collected images, the proportion of images with squatting, lying, standing, and walking gestures was 29%, 8%, 39%, and 24%, respectively. In this work, 2000 of these images were randomly chosen to generate the training dataset, which was used to train the human gesture detection model. The remaining 492 images were used as testing data to verify the performance of the proposed detection model.

B. IMAGE DATA ACQUISITION AUGMENTATION

Since different weather conditions, time of the day, seasons and other factors considerably influence the illumination, the generalization ability of the proposed human motion detection model depends on the integrity of the training dataset. To enhance the richness of the experimental dataset, the collected images were preprocessed in terms of the brightness, rotation, horizontal mirror, noise addition and blurring, as shown in Figure 6. After data augmentation, the numbers of images in the training and testing dataset increased to 16000 and 3936, respectively. The training and testing datasets contained 33264 labeled human gestures, with the numbers of "squatting", "lying", "standing" and "walking" being 7392, 8928, 9032 and 7912, respectively. The number of labeled "lying" and "standing" gestures were larger than that for the "walking" gesture, and the number of "squatting" gestures was the smallest. One image may contain multiple different human gesture targets and the completed dataset is shown in Table 1.

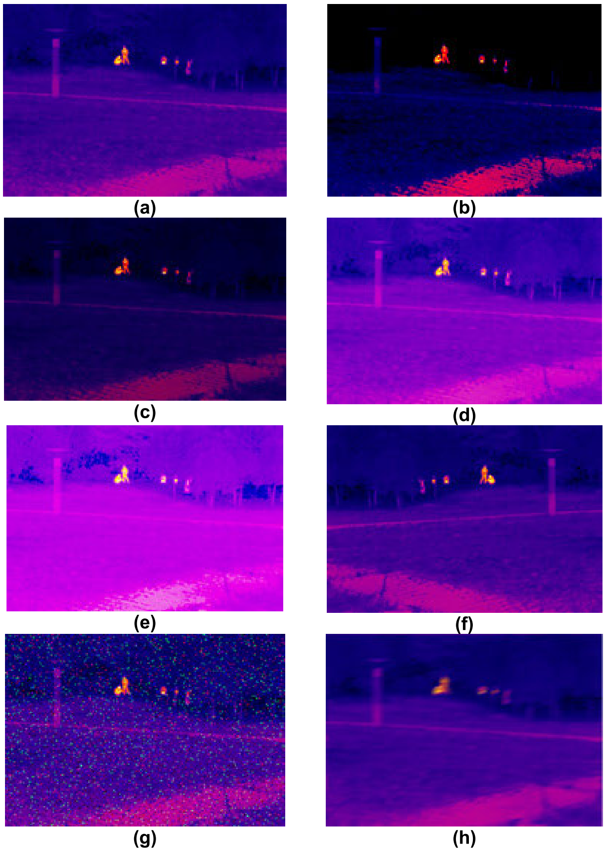


FIGURE 6. Image augmentation methods: (a) original image; (b–e) brightness transformation; (f) horizontal mirror; (g) salt and pepper noise addition; (h) blur processing.

IV. EXPERIMENT AND DISCUSSION

An image processing server with two NVIDIA 1080TI graphic cards was used to train and test the proposed human gesture detection model. The initialization parameters of the proposed network are listed in Table 2.

Considering the memory limit of the server, the input images were resized to a resolution of 416 × 416, and the batch was set as 16. We used 50,400 training steps to better analyze the training process. The initial learning rate was 0.001, and it was reduced to 0.0001 and 0.00001 after 30,000 and 45,000, respectively. The momentum of the network was 0.9, and the weight decay regularization was set as 0.0005. A series of testing experiments were conducted on

$$P\left(F_i \left| S_j(x)\right.\right) = \frac{S_i(x)\, P\left(S_j(x) \left| F_i\right.\right)}{S_i(x)\, P\left(S_j(x) \left| F_i\right.\right) + (1 - S_i(x))\, P\left(S_j(x) \left| B_i\right.\right)}$$

TABLE 1. The number of images generated by data augmentation methods.

Augmentation methods	Squatting	Lying	Standing	Walking	Total
Original data	723	199	972	598	**2492**
Brightness transformation	2892	796	3888	2392	**9968**
Horizontal mirror	723	199	972	598	**2492**
Noise addition	723	199	972	598	**2492**
Bur processing	723	199	972	598	**2492**
Total	**5784**	**1592**	**7776**	**4784**	**19936**

TABLE 2. Initialization parameters for the proposed network.

Size of image	Batch	Momentum	Initial learning rate	Decay	Training steps
416×416	16	0.9	0.001	0.0005	50400

the trained human gesture detection model by using testing images with a resolution of 640 × 480. The following indicators were used to evaluate the effectiveness of the human gesture detection model: precision and recall, *F1 score*, loss function, *IoU*, detection time, average precision (AP) and mean average precision (mAP), which have been widely used in the existing literature. Herein, we introduce the definition and function of these indicators.

The precision (P) and recall (R) can be defined as follows:

$$P = TP/(TP + FP); R = TP/(TP + FN)$$

where TP represents the true positive samples, FP denotes the false positive samples, TN represents the true negative samples, and FN denotes the false negative samples.

The AP represents the quality of the model in each category, and it can be obtained as

$$AP = \int_0^1 P(R)\, dR$$

The mAP indicates the quality of the model in all the categories, and it can be obtained as

$$AP = \left(\sum_{i=1}^{C} AP_i\right) \Big/ C$$

where C is the number of categories.

The F_1 score was used to evaluate the performance of the model, and it was determined as

$$F_1 = (2 \times P \times R)/(P + R)$$

In addition, the loss function was used to evaluate the performance of the network model, and it was determined as follows:

$$Loss = Error_{coord} + Error_{iou} + Error_{cls}$$

The coordinate prediction error $Error_{coord}$ can be expressed as

$$z^{Error_{coord}} = \lambda_{coord}\sum_{i=1}^{s^2}\sum_{j=1}^{B}L_{ij}^{obj}\left[\left(x_i-\hat{x}_i\right)^2+\left(y_i-\hat{y}_i\right)^2\right] + \lambda_{coord}\sum_{i=1}^{s^2}\sum_{j=1}^{B}L_{ij}^{obj}\left[\left(w_i-\hat{w}_i\right)^2+\left(h_i-\hat{h}_i\right)^2\right]$$

where λ_{coord} is the weight of the coordinate error; s^2 is the number of grids in the image; B is the number of bounding boxes generated by each grid; $L_{ij}^{obj} = 1$, if the object falls into the j_{th} bounding box in grid i and $L_{ij}^{obj} = 0$ otherwise; $\left[\hat{x}_i, \hat{y}_i, \hat{w}_i, \hat{h}_i\right]$ and $[x_i, y_i, w_i, h_i]$ denote the predicted and true values of the center coordinates, height, and weight of the predicted bounding box, respectively.

The error $Error_{iou}$ can be defined as follows:

$$Error_{iou} = \sum_{i=1}^{s^2}\sum_{j=1}^{B}L_{ij}^{obj}\left(C_i-\hat{C}_i\right)^2 + \lambda_{noobj}\sum_{i=1}^{s^2}\sum_{j=1}^{B}L_{ij}^{obj}\left(C_i-\hat{C}_i\right)^2$$

where λ_{noobj} is the weight of the IoU error; $\hat{C}_i$ and C_i denote the predicted confidence and true confidence, respectively.

The classification error $Error_{cls}$ can be defined as follows:

$$Error_{cls} = \sum_{i=1}^{s^2}\sum_{j=1}^{B}L_{ij}^{obj}\sum_{c\in classes}\left(p_i(c)-\hat{p}_i(c)\right)^2$$

where c is the class that the target belongs to. $p_i(c)$ and $\hat{p}_i(c)$ respectively refer to the true and predicted probability that the object belonging to class c lies in grid i.

The IoU is another criterion used to evaluate the detection accuracy, and it can be obtained by calculating the overlap ratio between the predicted and true bounding boxes, as follows:

$$IoU = S_{overlap}/S_{union}$$

where $S_{overlap}$ is the intersection area of the predicted and true bounding boxes, and S_{union} denotes the union area of the predicted and true bounding boxes.

The average detection times were also compared, as reported in this paper.

A. EFFECT OF DATA CATEGORIES

To verify the effect of the different categories in the dataset on the detection results, the infrared images with squatting, lying, standing and walking human gestures, were used to train and test the proposed human gesture detection neural network. The P–R curves of different categories for the proposed model were as shown in Figure 7.

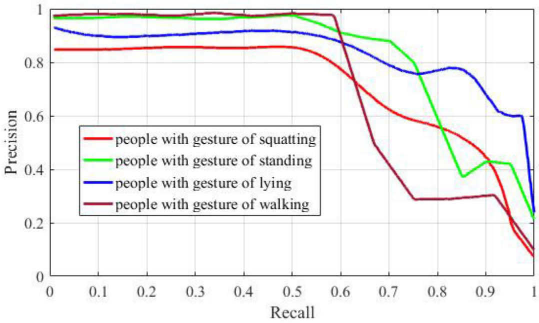


FIGURE 7. P-R curves of different categories for the proposed human gesture detection model.

TABLE 3. AP values of different data categories.

Class	F1 score	AP
"Squatting"	0.816	0.7054
"Lying"	0.897	0.7054
"Standing"	0.903	0.8279
"Walking"	0.832	0.7251
Average value	**0.862**	**0.7693**

TABLE 4. Classification prediction results (%).

Category	Squatting	*Lying*	*Standing*	*Walking*
Squatting	97.28	2.72	0	0
Lying	3.11	96.89		0
Standing	0	0	93.72	6.28
Walking	0	0	5.84	94.16

Mathematically, the AP is defined as the area under the P–R curve, reflecting the average performance of the algorithm under different IoU thresholds, and it was set as 0.5 in this work. The AP and F1 scores of different data categories are presented in Table 3.

The mAP is the mean of the AP values in a subclass, and its value was 76.93% for the proposed human gesture detection model.

To observe the boundary boxes, the "squatting", "lying", "standing" and "walking" categories were used to label the human gestures, as shown in Figure 8.

We considered the confusion matrix of the classification prediction results to evaluate the performance of the proposed method, as shown in Table 4. The values in the main diagonal denote the percentages of the correctly classified categories, and the remaining values correspond to the percentages of the incorrectly classified categories. It was noted that the main errors occurred when "squatting" was classified as "lying", "lying" was classified as "squatting", "standing" was classified as "walking", and "walking" was classified as "standing". We believe that "squatting" and "lying", as well as "standing" and "walking", are considerably similar in terms of the feature information. Furthermore, the sizes of the datasets are relatively small, and the datasets are relatively unbalanced. These aspects need to be further considered to solve the problem of interest.

From the above training and testing results, it can be noted that the categories of the targets affect the detection results of the proposed network. The number of humans with squatting gestures is relatively small, and thus, the detection results of human targets with squatting gestures are worse than those for human targets with other gestures. The proposed model obtained the best detection results for infrared images with walking human gestures due to the more notable features in the infrared images. This finding occurs because the body temperature rises to different degrees after walking and running motions.

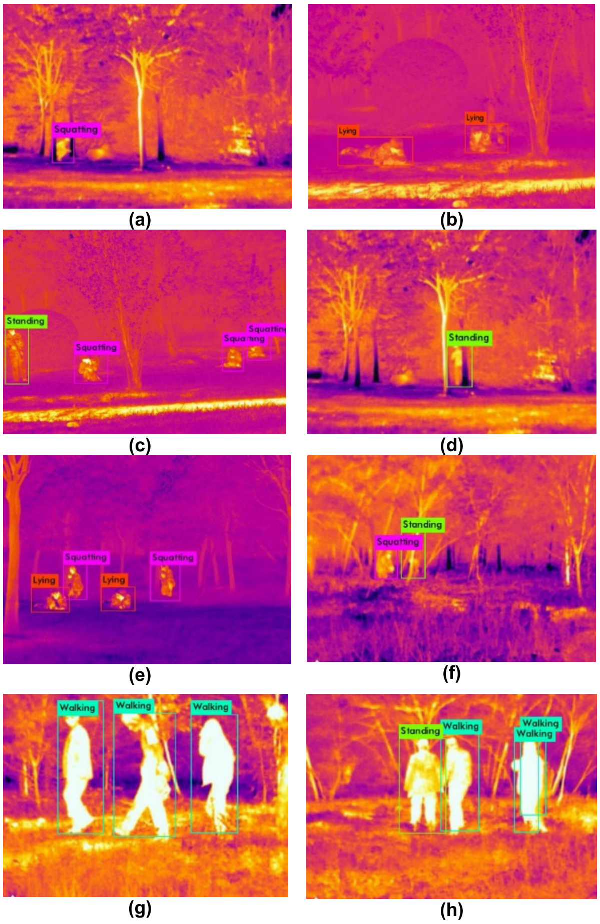


FIGURE 8. Detected humans with different gestures in the infrared images: (a) squatting; (b) lying; (c) standing; (d) walking; (e–i) combination of several gestures.

TABLE 5. Comparison of *F1*, *IoU*, mAP and average detection time.

Model	F1 score	IoU	mAP (%)	Time(s)
SSD	0.738	0.805	71.15	0.109
YOLO-V3	0.814	0.864	73.64	0.118
RetinaNet	0.846	0.857	74.79	0.104
YOLACT	0.846	0.868	75.23	0.113
Faster R-CNN	0.895	0.896	78.62	0.968
Proposed	0.862	0.873	76.93	0.122

B. COMPARISON WITH DIFFERENT DETECTION MODELS

The detection performances of the proposed model was compared with that of several other detection models, including

TABLE 6. **Classification results for the proposed network model when using different types of images.**

Data	F1 score	IoU	mAP (%)	Average time(s)
RGB images	0.664	0.759	66.17	0.118
Infrared images	0.773	0.838	72.32	0.107
Saliency infrared image pairs	0.862	0.873	76.93	0.122

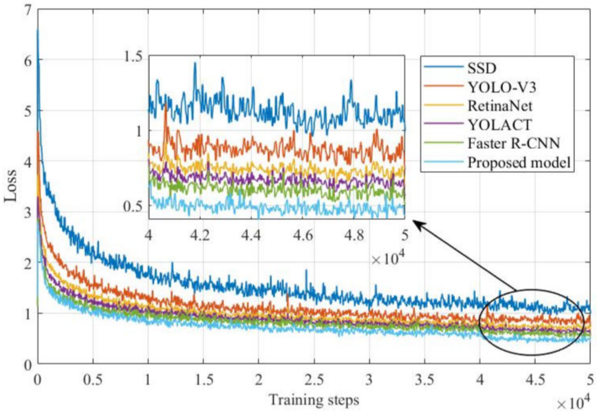


FIGURE 9. **Loss curves of several models.**

the single shot multibox detector (SSD), original YOLO-V3 RetinaNet [32], YOLACT [33] and faster R-CNN, to verify the superiority of the proposed human gesture detection model. The loss function curves of these detection models during training are as shown in Figure 9. The training results of different target detection models indicate that the average loss continuously reduces with an increase in the number of iterations, and the proposed detection model consistently exhibits a faster convergence in the training process. The final loss of the SSD, original YOLO-V3, RetinaNet, YOLACT, faster R-CNN and proposed detection model is approximately 1.21, 0.84, 0.72, 0.69, 0.62 and 0.51, respectively. In addition, compared to other target detection models, the loss curve of our proposed method exhibits a continuously decreasing trend during the training process until after 45000 training steps. These results demonstrates the better training performance of the proposed human gesture detection model.

The P–R curves for the SSD, original YOLO-V3, RetinaNet, YOLACT, faster R-CNN and the proposed model are as shown in Figure 10. The detection results pertaining to the *F1 score*s, *IoU* function *mAP* and average detection time of the target detection models are summarized in Table 5.

The F1 score and *IoU* value for the proposed network are approximately 0.862 and 0.873, respectively, which are higher than those for the SSD model, original YOLO-V3, RetinaNet and YOLACT models. Although the F1 score and *IoU* value of the faster R-CNN model are slightly higher than those of the proposed model (by 0.013 and 0.023, respectively), the average detection time is 0.968 s, which is approximately 8 times larger than that for the proposed model. This analysis indicates that the proposed model exhibits excellent processing speed performance while ensuring a high detection accuracy.

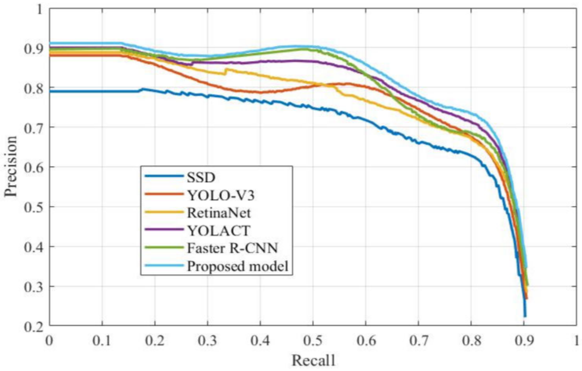


FIGURE 10. **P–R curves of several models.**

C. COMPARISON OF CLASSIFICATION RESULTS

The time span of the data collection process is from 9 a.m. to 8 p.m., and the light intensity varies considerably (from less than 50 lux to more than 50,000 lux). It is difficult to identify the gestures of the humans from the RGB images under the conditions of weak light intensity, especially in the evening and night time. To further verify the effectiveness of the proposed method, we trained and tested the proposed neural network model using pure RGB images, pure infrared images, and saliency thermal image pairs. Several RGB image samples with less observable human gestures and the corresponding infrared images and saliency infrared image pairs are shown in Figure 11.

As shown in Figure 11, the gestures of humans cannot be easily recognized in many cases when using an RGB image, such as in the presence of a street light under dusk conditions, street light in the evening, absence of street light in the late evening, occlusion and the human and background exhibiting similar colors. However, in these cases, the contour and brightness features of the humans are relatively more notable in the infrared image and the corresponding saliency images. The classification results for the proposed network model, as obtained using different types of images were compared, as presented in Table 6.

The classification results presented in Table 6 confirm that by using the saliency infrared image pairs and data fusion method, we can effectively improve the detection accuracy of the proposed network while maintaining a reasonable detection time.

D. DETECTION RESULTS UNDER THE CONDITION OF OCCLUSION AND OVERLAP

In outdoor scenes, the presence of occlusion due to trees, buildings and other structures, and the overlap between humans can affect the detection accuracy. The results of the F1 scores, *IoU* and AP for the proposed human gesture

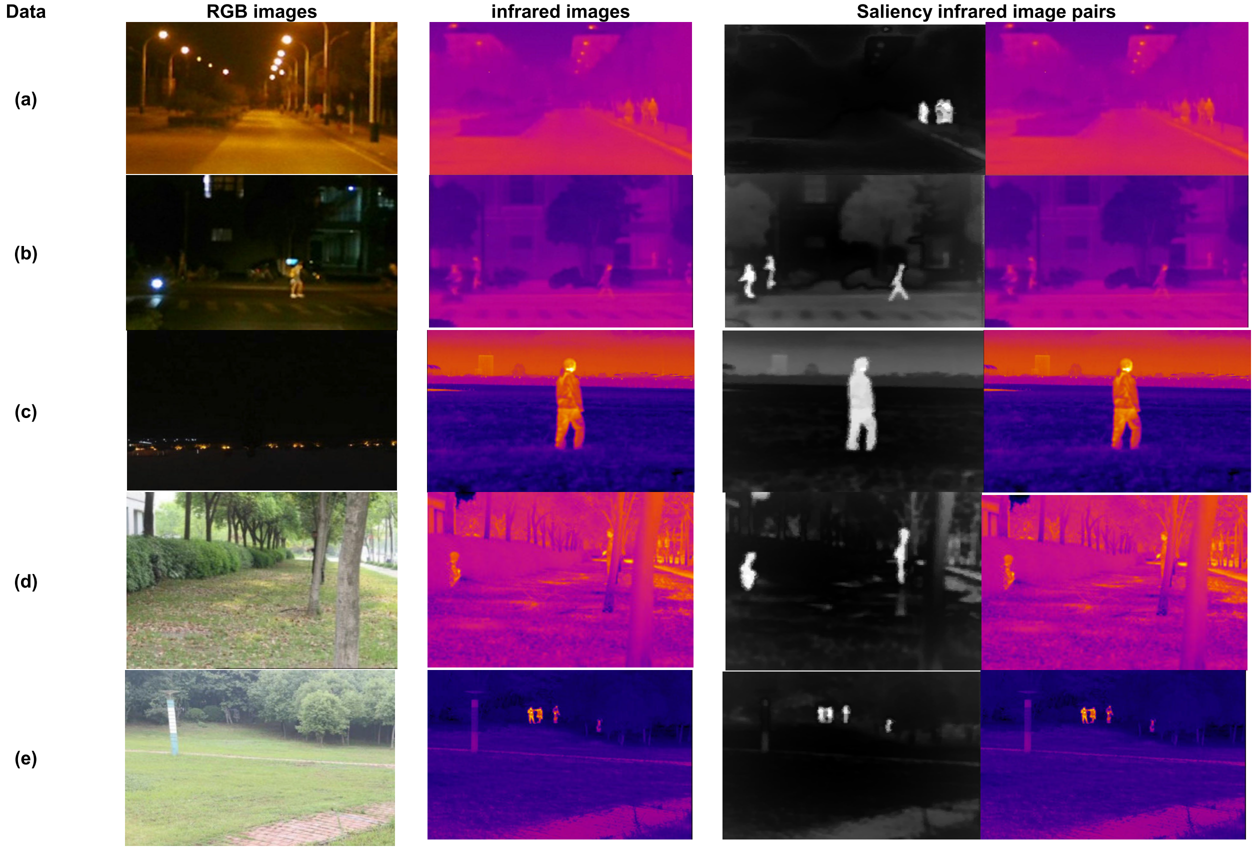


FIGURE 11. Pure RGB, pure infrared, and saliency thermal image pair dataset: (a) street light in dusk conditions; (b) dim street light in the evening; (c) condition without street light in the late evening; (d) occlusion; (e) human and background exhibit similar colors.

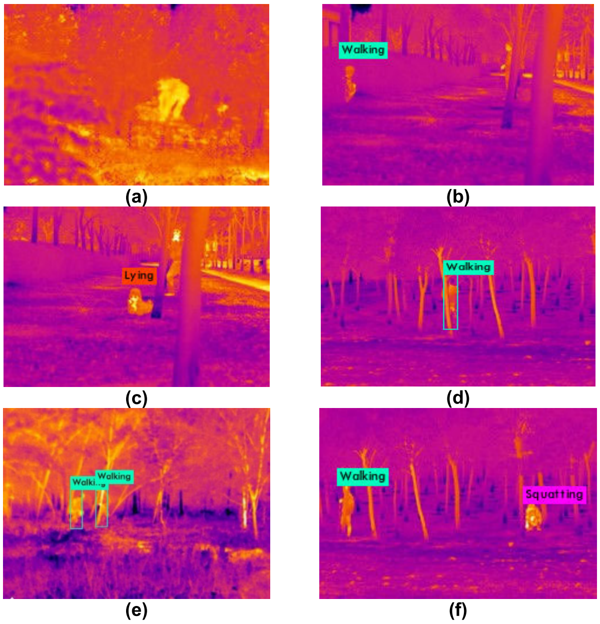


FIGURE 12. Missed and mistaken detection results of humans, owing to occlusion and overlap: (a–c) missed; (d–f) mistaken.

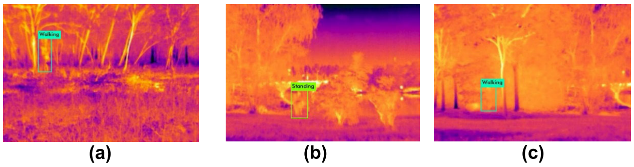


FIGURE 13. Incorrect recognition of humans with confidence values of (a) 53%; (b) 64%; (c) 51%.

detection model under the conditions of occlusion and overlap are shown in Fig. 12 and Table 7.

These detection results indicate that under occlusion and overlap conditions, the accuracy of the human gesture detection models is reduced. However, in most cases of occlusion and overlap, the proposed model still exhibits a satisfactory performance for human gesture detection in infrared images.

E. DETECTION RESULTS IN SCENES WITHOUT HUMANS

In an outdoor scene, it is possible for an infrared camera to capture images that do not contain human targets. We used 50 infrared images that did not contain human targets to verify the performance of the proposed detection model and to test whether the model would identify some humanoid targets as humans. Specifically, infrared images containing backgrounds of the sky, grass and buildings were collected. The detection results indicated that some humanoid branches were recognized as humans, as shown in Figure 13.

These test results indicate that the proposed detection model can detect most of the human targets in infrared images, even under some severe occlusion and overlap

TABLE 7. F1 scores, IoU and AP values under occlusion and overlap conditions.

Class	F1 score	IoU values	AP
"Squatting"	0.726	0.784	0.6298
"Lying"	0.795	0.812	0.7311
"Standing"	0.803	0.829	0.7392
"Walking"	0.732	0.789	0.6474

conditions. However, in some scenarios, humanoid targets may still be identified as humans. This problem can likely be solved by increasing the scale of the target dataset under a larger number of scenarios and environmental conditions.

V. CONCLUSION

This work reports upon a deep learning approach for human gesture detection in infrared images, based on the improved YOLO-V3 network. The proposed model uses three DenseNet blocks, added before the residual components in the YOLO-V3 network, to enhance the convolutional feature propagation and improve the human gesture detection performance. The saliency maps of the infrared images were detected as an additional input channel for the network to improve the robustness and performance of the proposed human gesture detection model. To verify the detection performance of the proposed model, several experiments were conducted, and the results indicated that the proposed network has a better detection performance than those of the original detection network YOLO-V3 and SSD. The detection accuracy of the proposed method is comparable to that of the faster R-CNN, which is a state of the art network in terms of the accuracy. However, the proposed network exhibits a notable advantage in terms of the detection time performance. The proposed model is capable of human gesture detection under low visibility images, such as in rainy and foggy weather, night time conditions, and conditions in which the colors of the targets and the background are similar. In addition, the proposed model exhibits a high performance for human gesture detection under conditions involving occlusion and overlap. In the future, we aim to further optimize the human gesture dataset of the infrared images and predict the dynamic behavior of humans.

REFERENCES

[1] R. Hemati and S. Mirzakuchaki, "Using local-based Harris-PHOG features in a combination framework for human action recognition," *Arabian J. Sci. Eng.*, vol. 39, no. 2, pp. 903–912, Feb. 2014.

[2] D. Carbonera Luvizon, H. Tabia, and D. Picard, "Learning features combination for human action recognition from skeleton sequences," *Pattern Recognit. Lett.*, vol. 99, pp. 13–20, Nov. 2017.

[3] F. Murtaza, M. H. Yousaf, and S. A. Velastin, "Multi-view human action recognition using histograms of oriented gradients (HOG) description of motion history images (MHIs)," in *Proc. 13th Int. Conf. Frontiers Inf. Technol. (FIT)*, Dec. 2015, pp. 297–302.

[4] G. M. Al, L. Zhang, and Y. Gotoh, "Spatio-temporal SIFT and its application to human gesture classification," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2012, pp. 301–310.

[5] T. S. Murray, D. R. Mendat, K. A. Sanni, P. O. Pouliquen, and A. G. Andreou, "Bio-inspired human action recognition with a micro-Doppler sonar system," *IEEE Access*, vol. 6, pp. 28388–28403, 2018.

[6] M. Faraki, M. Palhang, and C. Sanderson, "Log-Euclidean bag of words for human action recognition," *IET Comput. Vis.*, vol. 9, no. 3, pp. 331–339, Jun. 2015.

[7] D. Das Dawn and S. H. Shaikh, "A comprehensive survey of human action recognition with spatio-temporal interest point (STIP) detector," *Vis. Comput.*, vol. 32, no. 3, pp. 289–306, Mar. 2016.

[8] N. Li, X. Cheng, S. Zhang, and Z. Wu, "Realistic human action recognition by fast HOG3D and self-organization feature map," *Mach. Vis. Appl.*, vol. 25, no. 7, pp. 1793–1812, Oct. 2014.

[9] C. Chen, M. Liu, H. Liu, B. Zhang, J. Han, and N. Kehtarnavaz, "Multi-temporal depth motion maps-based local binary patterns for 3-D human action recognition," *IEEE Access*, vol. 5, pp. 22590–22604, 2017.

[10] T. Wang, Y. Chen, M. Zhang, J. Chen, and H. Snoussi, "Internal transfer learning for improving performance in human action recognition for small datasets," *IEEE Access*, vol. 5, pp. 17627–17633, 2017.

[11] H. J. Kim, J. S. Lee, and H. S. Yang, "Human gesture recognition using a modified convolutional neural network," in *Proc. Int. Symp. Neural Netw.* Berlin, Germany: Springer, 2007, pp. 715–723.

[12] T. Minh Le, N. Inoue, and K. Shinoda, "A fine-to-coarse convolutional neural network for 3D human action recognition," 2018, *arXiv:1805.11790*. [Online]. Available: http://arxiv.org/abs/1805.11790

[13] Y. Wang, W. Zhou, Q. Zhang, and H. Li, "Visual attribute-augmented three-dimensional convolutional neural network for enhanced human action recognition," 2018, *arXiv:1805.02860*. [Online]. Available: http://arxiv.org/abs/1805.02860

[14] F. Meng, H. Liu, Y. Liang, M. Liu, and W. Liu, "Hierarchical dropped convolutional neural network for speed insensitive human action recognition," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2018, pp. 1–6.

[15] B. Meng, X. Liu, and X. Wang, "Human action recognition based on quaternion spatial-temporal convolutional neural network and LSTM in RGB videos," *Multimedia Tools Appl.*, vol. 77, no. 20, pp. 26901–26918, Oct. 2018.

[16] H. Yang, C. Yuan, J. Xing, and W. Hu, "SCNN: Sequential convolutional neural network for human action recognition in videos," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 355–359.

[17] C. Li, S. Sun, X. Min, W. Lin, B. Nie, and X. Zhang, "End-to-end learning of deep convolutional neural network for 3D human action recognition," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops (ICMEW)*, Jul. 2017, pp. 609–612.

[18] S. Ji, W. Xu, and M. Yang, "3D convolutional neural networks for human gesture recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 221–231, 2013.

[19] L. Sun, K. Jia, D.-Y. Yeung, and B. E. Shi, "Human action recognition using factorized spatio-temporal convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4597–4605.

[20] P. Bhattacharjee and S. Das, "Two-stream convolutional network with multi-level feature fusion for categorization of human action from videos," *Pattern Recognit. Mach. Intell.*, vol. 10597, pp. 549–556, Dec. 2017.

[21] A. Akula, A. K. Shah, and R. Ghosh, "Deep learning approach for human gesture recognition in infrared images," *Cogn. Syst. Res.*, vol. 50, pp. 146–154, Jan. 2018.

[22] J. F. Li and W. G. Gong, "Application of thermal infrared imagery in human action recognition," *Adv. Mater. Res.*, vols. 121–122, pp. 368–372, Jun. 2010.

[23] H. Osada, S. Chiba, H. Oka, and K. Seki, "Human action pattern monitor for telecare system utilizing magnetic thin film infrared sensor," *J. Magn. Magn. Mater.*, vol. 239, nos. 1–3, pp. 576–578, Feb. 2002.

[24] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[25] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.

[26] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*. [Online]. Available: http://arxiv.org/abs/1804.02767

[27] K. Jo, J. Im, J. Kim, and D.-S. Kim, "A real-time multi-class multi-object tracker using YOLOv2," in *Proc. IEEE Int. Conf. Signal Image Process. Appl. (ICSIPA)*, Sep. 2017, pp. 507–511.

[28] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot MultiBox detector," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 21–37.

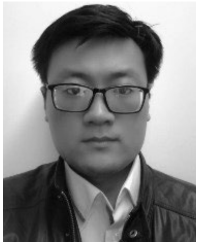
[29] H. Hu, D. Dey, A. Del Giorno, M. Hebert, and J. Andrew Bagnell, "Log-DenseNet: How to sparsify a DenseNet," 2017, *arXiv:1711.00002*. [Online]. Available: http://arxiv.org/abs/1711.00002

[30] X. Ren and J. Malik, "Learning a classification model for segmentation," in *Proc. 9th IEEE Int. Conf. Comput. Vis.*, vol. 1, Oct. 2003, pp. 10–17.

[31] M. Ester, H. P. Kriegel, and J. Sander, "A density-based algorithm for discovering clusters in large spatial databases with noise," *KDD*, vol. 96, no. 34, pp. 226–231, 1996.

[32] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 318–327, Feb. 2020.

[33] D. Bolya, C. Zhou, F. Xiao, and Y. J. Lee, "YOLACT: Real-time instance segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9157–9166.



KEKE GENG received the B.S. degree in mechanical engineering from the Nanjing University of Technology and Engineering, Nanjing, China, in 2010, and the M.S. and Ph.D. degrees in system control and information processing from the Department of Information and Control Systems, Bauman Moscow State Technical University, Moscow, Russia, in 2013 and 2017, respectively. He is currently an Assistant Professor with the School of Mechanical Engineering, Southeast University, Nanjing. His current research interests include the autonomous motion control of vehicles, intelligent environmental awareness, integrated navigation systems, and stability of complex control systems.

GUODONG YIN (Senior Member, IEEE) received the Ph.D. degree in vehicle engineering from Southeast University, Nanjing, China, in 2007. From 2011 to 2012, he was a Visiting Research Scholar with the Department of Mechanical and Aerospace Engineering, Ohio State University, Columbus, OH, USA. He is currently a Professor with the School of Mechanical Engineering, Southeast University. His current research interests include vehicle dynamics and control, connected vehicles, and multiagent control.