Thermal Infrared Human Recognition Based on Multi-scale Monogenic Signal Representation and Deep Learning

Yong Tan, Wenjuan Yan, Shijian Huang, Derong Du, Liangping Xia

***Abstract*—the infrared human recognition is challengeable due to the factors including poor imaging quality, disturbance objects in surroundings, large variations of human poses and casual movements. In this paper, a novel human recognition method is proposed. Its critical components include a feature descriptor that is referred to as a histogram of oriented monogenic energy (HOME), and a deep learning network that is referred to as a deep brief network (DBN). The feature descriptor, which is formulated from the multi-scale monogenic signal representation (MMSR), provides discriminative representation of lines/edges of the human subjects of interest. The DBN learns multiple layers of abstraction of the feature and conducts accurate human and/or non-human classification. Experimental results validate the advantages of the proposed method in recognition accuracy and robustness to scenic changes as well, due to such factors including the discriminative representation of human cues, high-level understanding of the cues, and tightly architectural coupling between the feature and the classifier.**

***Index Terms*—infrared human recognition, multi-scale monogenic signal representation, histogram of oriented monogenic energy, deep brief network, contrastive divergence, maximum likelihood estimator**

I. INTRODUCTION

In recent years, thermal infrared imagery has been extensively applied to such application as all-weather surveillance, traffic monitoring and behaviour analysis due to its ability to work under poor lighting conditions. In the referred applications, human subjects are often the most important objects of interest because they are the major participants in related events of interest. Recognition of the visually captured human subjects is important as it is the fundamental step for many higher level functions. However, infrared human recognition is a challengeable task [1] [2]. First, the quality of infrared imaging is generally poor, owing to the factors including blurred boundaries, loss of colour and absence of textures. Second, disturbance objects with similar brightness or shapes to human subjects may degrade the recognition accuracy. Third, modelling human subjects with considerable variations in size, appearance, and pose is difficult. Fourth, describing human movements in a relatively unified way is challenging because of their casualness. To address the difficulties, a human recognition method consisting of a feature description referred to as a histogram of oriented monogenic energy (HOME) and a deep learning network as a deep brief network (DBN) [3] is proposed in this paper. The HOME originates from the multiscale monogenic signal representation (MMSR) [4] [5] and provides a compact representation of human subjects. The DBN conducts higher-level abstraction of the representation and provides accurate human/non-human prediction. As a whole, the method enjoys advantages in recognition accuracy and robustness to scenic changes, which make it practical for human recognition systems.

The remainder of this paper is organized as follows. Related studies of infrared human recognition are discussed in section II. The MMSR and the restricted Boltzmann machine (RBM), both of which are the preliminary fundamentals of the proposed method, are introduced in section III. In section IV, details of the proposed method are explained. In section V, experimental results and an algorithmic analysis are provided. At last, concluding remarks are presented in section VI.

Manuscript received September 18, 2019; revised February 3, 2020. This work was supported by the Scientific and Technological Research Program of Chongqing Municipal Education Commission under Grant KJZD-K201901402.

Yong Tan and the co-authors are all with the School of Electronic Engineering, Yangtze Normal University, Fuling district, Chongqing, 408003, China. Yong Tan is the corresponding author and his e-mail address is cquty@126.com.

II. RELATED WORKS

The methods proposed for infrared human recognition include human body modelling, template matching and statistical classification. Human body modelling methods [6] [7] model human shapes using outlines, skeletons and body parts, and therefore facilitate the analysis of human poses and/or actions. However, these methods often fail to detect accurate human shapes due to occlusions and complex backgrounds. Template matching methods [8] [9] calculate the similarity between a candidate and each of the samples that belong to a template set defining comprehensive or partial human poses. Major difficulties are the enormous calculation cost and the construction of a compact and representative set of human templates. Statistical classification methods are popular due to their adaptation to great human variations. Generally, a statistical classification method consists of a learning stage and an inference stage. In the learning stage, a specific classifier is learned from a series of features derived from the dataset composed of a series of human samples and non-human samples. In the inference stage, a well-trained classifier conducts pattern classification, in which a candidate is inferred to be a human subject or a non-human subject.

Discriminative feature description and highly generalized classification contribute to the success of a statistical

classification method. With respect to feature descriptors, the histogram of oriented gradient (HOG) [10] and its invariants [2] [11] [12] may be the most popular. The HOG feature family often provides acceptable recognition performance but may fail with subjects with unclear contours caused by flat gradients or strong noise. Compared with the HOG family, such histogram-based features as the coded similarity histogram of local intensity differences (CSHLID) [13] and shape distribution histogram (SDH) [14] provide more stable and tractable representation of image edges, due to their robustness to weak edges and noise. However, they may still be degraded by environmental objects. Point-based features, such as multi-block local binary pattern (LBP) [15], centre-symmetric LBP [16] and codebook of SURF [17], can be discriminative but may not perform well if they encounter large subject motions or bad choices of key points. Multiple features that originate from different sensors have also been proposed [18] [19] [20]. Fusion of multiple features does improve recognition performance at the cost of increasing computations if the feature sources are truly complementary. For classification, many classifiers, such as the support vector machine (SVM) with linear or nonlinear kernels [2] [10] [11] [17], boosting [12] [15], modified sparse representation classification (MSRC) [14], random forest [16] and relevance vector machine (RVM) [21], have been proposed for human/non-human classification, and they show both advantages and disadvantages as well. Recently, deep neural nets, which are part of the broader machine learning field of learning representations of data, have attracted attention because of their ability of pattern analysis (unsupervised) and classification (supervised) that outperforms many rivals. Some trials of applying deep nets [20] [22] to human recognition tasks have already achieved inspiring performance when the nets have been properly trained. However, deep neural nets generally require large-scale datasets with clean annotations and excessive amount of training time. Despite considerable progress, infrared human recognition techniques need to be further explored due to insufficient performance in complex environments.

III. Preliminary Fundamentals

A. The MMSR Fundamentals

The monogenic signal is built around the Riesz transform. Let $f(z)$ with $z=(x,y)$ be a 2D signal. Its Riesz transform is defined as

$$f_R(z)=(f_x(z),f_y(z))=(h_x * f(z),h_y * f(z)) \quad (1)$$

where (h_x,h_y) represents the spatial representation of the Riesz kernel, namely

$$(h_x,h_y)=\left(\frac{x}{2\pi\|z\|^3},\frac{y}{2\pi\|z\|^3}\right) \quad (2)$$

Combining $f(z)$ with $f_R(z)$ produces the monogenic signal denoted by

$$f_M(z)=(f(z),f_x(z),f_y(z)) \quad (3)$$

where $f(z)$ is the real part, and $f_x(z),f_y(z)$ are the imagery parts. From $f_M(z)$, three orthogonal components, including the local amplitude A, local phase ϕ and local orientation θ, which respectively describe local energetic information, local structural information and geometric information, can be calculated as follows

$$(A,\phi,\theta)=\left(\sqrt{f^2+f_x^2+f_y^2},-sign(f_x)a\tan 2\left(\sqrt{f_x^2+f_y^2}/f\right),a\tan\left(f_y/f_x\right)\right) \quad (4)$$

Assume that $f(z)$ refers to a real digital image, which has a finite size and is composed of an extensive range of frequencies. The image needs to be processed by band-pass filtering before applying the Riesz transform. Among band-pass filters, the log-Gabor families are rather advantageous due to three reasons. First, they offer the best simultaneous spatial and frequency localization. Second, they avoid the bandwidth limitation from which traditional Gabor filters suffer. Third, they enable desirable high-pass characteristics due to their Gaussian-shaped response along the logarithmic frequency scale.

Let $G(\omega)$ be the frequency response of the log-Gabor filter expressed as

$$G(\omega)=\exp\left\{-[\log(\omega/\omega_0)]^2/[\log(\sigma/\omega_0)]\right\} \quad (5)$$

where ω_0 is the centre frequency, and σ is the scaling factor of the bandwidth. As suggested, the ratio $\sigma_r=\sigma/\omega_0$ can be a constant. Now, the monogenic signal representation can be rewritten as

$$f_{lg-M}(z)=(f_{lg}(z),h_x * f_{lg}(z),h_y * f_{lg}(z)) \quad (6)$$

where $f_{lg}(z)=f(z)*F^{-1}(G(\omega))$. Note F^{-1} denotes a 2D inverse Fourier transform.

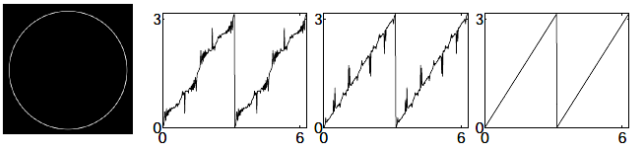
By varying the scale index k of $G(\omega)$, MMSR, which provides a full description of $f(z)$ can be obtained. The corresponding parameters σ and ω_0 can be respectively determined by $\omega_0=(\lambda_{\min}\mu^{s-1})^{-1}$ and $\sigma=\sigma_r\omega_0$, where $\lambda_{\min}$ is the minimal wavelength, μ is the multiplication factor of the wavelength and σ_r is a ratio constant. By using this filter bank, it reaches the multi-scale representation of a monogenic signal $\{f_M^1,f_M^2,...,f_M^s\}$ and the corresponding multi-scale monogenic components

$$\{A_1,\phi_1,\theta_1,A_2,\phi_2,\theta_2,...A_k,\phi_k,\theta_k,...,A_s,\phi_s,\theta_s\} \quad (7)$$

where A_k,ϕ_k and θ_k, $k=1,2,...,S$ are the local amplitude, local phase and local orientation of the monogenic signal at the k th scale, respectively.

The MMSR provides a reasonable description of the local properties of an image. First, the monogenic signal presents an appropriate local signal representation that preserves the split of its identity based on three local components. Second, the MMSR causes considerably less representation redundancy. The redundancy in the MMSR only originates from a multi-scale representation due to the orthogonality among the monogenic components on an arbitrary scale. For such rivals as the multi-scale Gabor wavelet representation, not only redundancy across multiple scales but also redundancy across multiple orientations exists. Third, the monogenic components provide rather meaningful information about local regions. In detail, the local amplitude describes the local energetic information. The relatively high amplitude generally indicates higher energetic local features, including edges, lines, and textures. The local orientation reveals important geometrical information, which includes the dominant orientation of local structures. As in Fig. 1, the MMSR does reach the desirable orientation of the circle

image which must be a (piecewise) linear function, while its rivals such as Sobel operator and spherical quadrature filter (SQF) fail. With respect to the local phase, the MMSR indicates information about the local structure, such as step edges and spots along the dominant orientations.



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

Fig. 1. Orientation estimation of the (b) optimized Sobel operator, (c) SQF and (d) MMSR on (a) the circle image.

B. The RBM Fundamentals

A RBM consists of a layer of visible units that represent the data and a layer of hidden units that learn to represent the feature that capture higher-order correlations in the data. The visible layer and the hidden layer are connected by a matrix of symmetrically weighted connections without internal connections within each layer.

An overview of the RBM can be seen in Fig. 2. A small circle indicates a neuron unit, and a line that connects two circles represents the weighted connections between two neuron units. The variable $\boldsymbol{v} = (v_1, v_2, ..., v_m)^T$ denotes the unit values in the visible layer, $\boldsymbol{h} = (h_1, h_2, ..., h_n)^T$ represents the unit values in the hidden layer, $\boldsymbol{a} = (a_1, a_2, ..., a_m)^T$ indicates the neuron biases in the visible layer, $\boldsymbol{b} = (b_1, b_2, ..., b_n)^T$ denotes the neuron biases in the hidden layer, and $W = \{w_{ij}\} \in R^{m \times n}$ represents the weight matrix between the visible layer and the hidden layer.

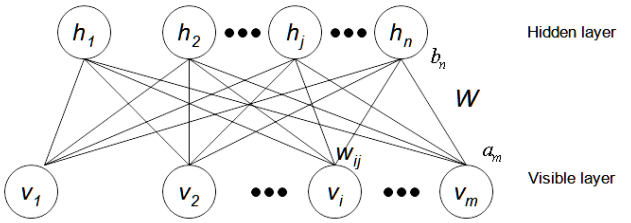


Fig. 2. An overview of the RBM.

Given ($\boldsymbol{v}$, $\boldsymbol{h}$), the RBM energy can be calculated by

$$E(v,h) = -\sum_{i=1}^{m} a_i v_i - \sum_{j=1}^{n} b_j h_i - \sum_{i=1}^{m}\sum_{j=1}^{n} w_{ij} v_i h_i \tag{8}$$

Note a more desirable RBM should have lower energy.

The conjunctive probability of $\boldsymbol{v}$ and $\boldsymbol{h}$ can be calculated by

$$P(v,h) = \frac{\exp[-E(v,h)]}{\sum_{v}\sum_{h}\exp[-E(v,h)]} \tag{9}$$

now, the probability of the visible units can be calculated by

$$P(v) = \sum_{h} P(v,h) \tag{10}$$

The contrastive divergence (CD) method [23] is generally employed for RBM training that operates for $P(v)$. Its rules for updating the network parameters are as follows

$$\begin{cases} \Delta w_{ij} = \varepsilon\left(\langle v_i h_j \rangle_{data} - \langle v_i h_j \rangle_{recon}\right) \\ \Delta a_i = \varepsilon\left(\langle v_i \rangle_{data} - \langle v_i \rangle_{recon}\right) \\ \Delta b_j = \varepsilon\left(\langle h_j \rangle_{data} - \langle h_j \rangle_{recon}\right) \end{cases} \tag{11}$$

where $\langle ... \rangle_{data}$ represents an expectation with respect to the distribution of the training data, $\langle ... \rangle_{recon}$ denotes an expectation with respect to the CD approximation of the model distribution, and ε is the learning rate. After training, the hidden unit values are considered as the feature extracted from the visible unit values.

IV. The Proposed Method

A diagram of the proposed method is shown in Fig. 3. At the beginning, thermal infrared cameras are utilized to acquire infrared images and/or videos. Then, a robust segmentation method [25], which works on both image intensity and object motions, runs to detect saliency blobs that may contain human subjects. Next, candidate selection produces regions of interest (ROIs) by refining the blobs with prior knowledge about human appearances. Later, it generates the HOME feature from each ROI. At last, a trained DBN conducts supervised classification based on the HOME feature inputs. In this section, details of the HOME feature and the DBN are introduced.

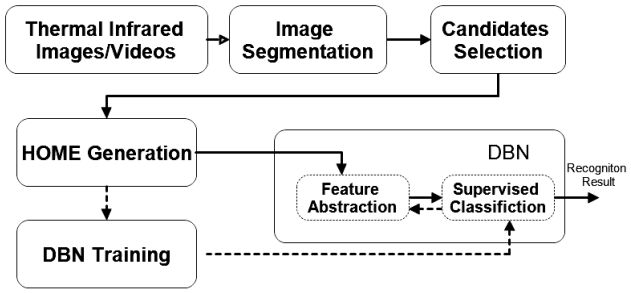


Fig. 3. Block diagram of the proposed method.

A. HOME Generation

The edges/lines in the infrared images provide the most discriminative cues for recognition of infrared objects with the absence of textures and colour. Unfortunately, the edges/lines are regularly blurred and polluted by imaging noise. Extraction of these discriminative cues from such structures remains challenging. As mentioned in section II, the monogenic components, especially the local amplitude and the local orientation, provide meaningful description of the edges/lines, since the local amplitude cannot be substantially affected by blurred image structures or noise, and the local orientation provides accurate estimation of the main orientation of the blurred and noisy edges/lines across a relatively larger image region. Therefore, based on the monogenic components, the HOME could be formulated to provide reasonable description of the degraded edges/lines, which primarily defines human subjects in infrared images.

Let f be an ROI. According to (7), the MMSR concerning about f can be calculated. With the available $A_k, k = 1, 2, ..., S$ and $\theta_k(x, y), k = 1, 2, ..., S$, the feature HOME can be generated through the following steps.

Step 1. Divide each $A_k, k=1,2,...,S$ and each $\theta_k(x,y), k=1,2,...,S$ maps into $P \times Q$ non-overlapping cells.

Step 2. Divide orientation range $[0, \pi]$ into N directional bins. Unify $\theta_k(x,y), k=1,2,...,S$ into these bins and therefore obtain the corresponding bin index $i, i \in [1, N]$. By replacing $\theta_k(x,y)$ with the bin index i, the discrete orientation map $\theta_k^{'}, , k=1,2,...,S$ can be obtained.

Step 3. From $A_k, k=1,2,...,S$, calculate the directional energy histogram $h_j^k, j=1,2,...,P \times Q$ for the jth cell by summing all cell elements that have identical bin indices in $\theta_k^{'}, , k=1,2,...,S$ with respect to each of the N main directions.

Step 4. Scan A_k cells in a zigzag way and connect the corresponding energy histogram $h_j^k, j=1,2,...,P \times Q$ into $H_k, k=1,2,...,S$, namely, $H_k = \{h_1^k, h_2^k, ..., h_{P \times Q}^k\}$.

Step 5. By sequentially connecting $H_k, k=1,2,...,S$, it reaches the HOME $H = \{H_1, H_2, ..., H_S\}$.

Several exemplar images have been given in Fig. 4 in order to illustrate the formulation of H_k from the monogenic components on the k th scale.

B. DBN Construction

Deep neural nets (DNNs) are powerful to provide a hierarchical data representation that corresponds to different levels of abstraction via the (unsupervised) learning of multiple levels of features or representations of data [24]. The DBN, which is a representative DNN, is adopted for automatic feature abstraction and classification as well in the proposed method considering its flexibility in configuration and rapid training.

A DBN consists of a visible/input layer, several hidden layers and a label layer. With the exception of the label layer, which converts the learned representation into supervised predictions, two arbitrary adjacent layers construct an RBM. Thus, the DBN can be considered as several stacked RBMs plus a label layer.

In Fig. 5, it shows the proposed network architecture, in which VL, $HL^i, i=1,2,...,M$, and LA denote the visible layer, the hidden layers and the label layer, respectively. Since the VL works as the interface with the HOME feature, its unit number is set as the feature dimension. With respect to the hidden layers, the unit number of each layer d is set as the intrinsic dimension of HOME, or the least dimension that describes this input feature without a significant loss of cues. Note this parameter can be determined by the maximum likelihood estimator (MLE) [25]. In this way, it enables the higher-level abstraction of the feature as compact as possible and it assures the effectiveness of the greedy training algorithm for the DBN [26]. Unlikely, the number of hidden layers M, which determines the network depth, should be experimentally determined, because a deeper network depth potentially contributes to better classification due to the learned feature of higher-level abstraction but excessive depth degrades the training efficiency and generalization ability. As for the label layer, it adopts a logistic regression classifier. Its unit number is set to 2 as there are only two kinds of labels, i.e., humans or non-humans.

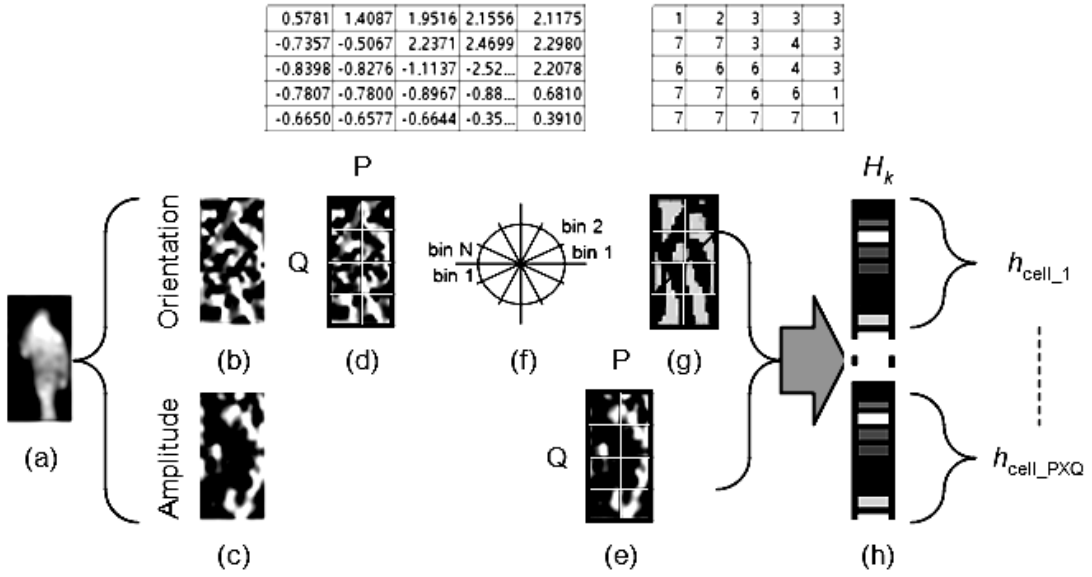


Fig.4. Exemplar images to illustrate the formulation of a feature vector from monogenic components on a specific scale. (a) The ROI; (b) θ_k on the k th scale; (c) A_k on the k th scale; (d) cell grids overlapped θ_k; (e) cell grids overlapped A_k; (f) setting of directional bins; (g) discrete orientation map; and (h) feature vector that corresponds to the k th scale.

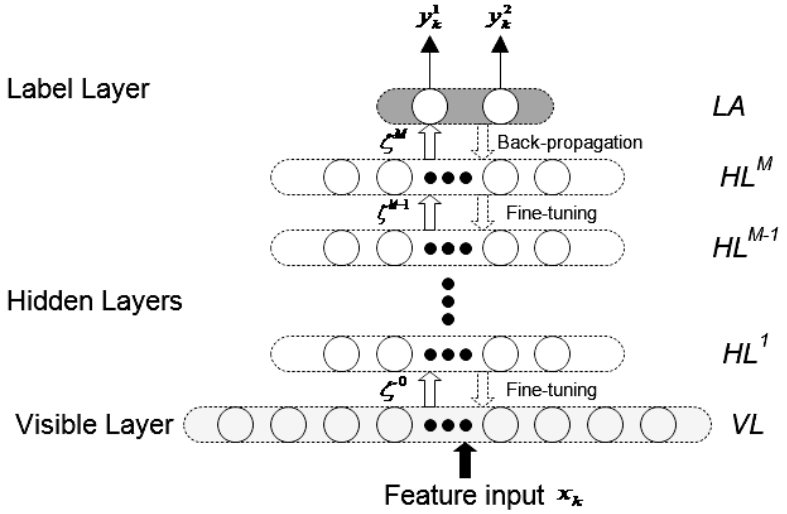


Fig. 5. Proposed architecture of the DBN.

Let X be the matrix that is regarded as a set of training samples. It is expressed as $X=[x_1,x_2,...,x_k,x_K]$, where x_k is a training sample, and K is the number of training samples. Let Y denote the labels that correspond to X. It is expressed as $Y=[y_1,y_2,...,y_k,y_K]$. Note that $y_k=[y_k^1,y_k^2]$ is the label vector. It is (1, 0) if x_k belongs to a human subject. Otherwise, it is (0, 1). Via a training process, a mapping function from the matrix input X to the label data Y can be learned over the network. With such function, unknown samples can be predicted as human or non-human subjects.

The training of a DBN consists of the pre-training and fine-tuning stages. Pre-training consecutively determines the parameters of two adjacent layers which can be viewed as a RBM by using the CD method. Fine-tuning refines the pre-trained layer parameters with the label layer information by using back propagation that attains the objective of minimizing the classification error $-\sum_t y_t \log y_t$, where y_t and y_t are the real label and the output label of data x_t on the M th layer. Let $\zeta^i=(W^i,a^i,b^i), i=0,1,...,M$ be layer-wise network parameter vectors.

The training procedure can be written as follows.

Pre-training stage

Step 1. Train an RBM on X to obtain ζ^0 between the lower two layers of the network.

Step 2. Transform X by the RBM to obtain data X' by sampling or computing the mean activation of the hidden units.

Step 3. Repeat this procedure with X←X' for the next pair of layers, and obtain ζ^i until the top two hidden layers have been attained.

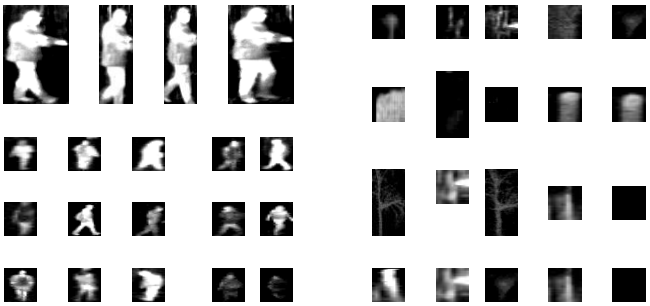
Fine-tuning stage

Step 4. Fine-tune $\zeta^i=(W^i,a^i,b^i), i=0,1,...,M$ using the backpropagation and then find the local optimum parameters $\zeta^{i*}, i=0,1,...,M$.

This procedure is also illustrated in Fig. 5. A properly trained DBN is ready for incoming human recognition.

V. Experiment Results and Algorithmic Analysis

The proposed method (PRO) has been implemented in MATLAB on a personal computer with an Intel Core I7 4720HQ 2.60 GHz CPU, 4G RAM, and Windows 10 operating system. A series of infrared sequences are selected from the OTCBVS database [27] and the custom databases shot by an FLIR A40 thermal infrared camera with a resolution of 320×240 pixels for test. In Table I, the characteristics of three representative sequences that differ in human size, pose, motion pattern, and imaging quality are listed. From the sequences, a dataset having 3,000 human blobs and 3,000 non-human blobs is firstly extracted and then employed to verify the PRO. Parts of the samples are shown in Fig. 6.



(a) Human samples (b) Non-human samples

Fig. 6. Representative human/non-human samples.

With respect to the DBN parameters, the bias for visible units and hidden units were initially set to 0. The learning rate ε and the epochs were respectively fixed to 0.01 and 30. The weights for the neuron connections were randomly initialized between [0, 1], which satisfy the Gaussian distribution. The number of hidden layers was set to 2, which rendered the network architecture as simple as possible. The unit number of each hidden layer d was set to the MLE estimated intrinsic dimension of the input feature, so that it couples the network architecture with the network input. To objectively evaluate the recognition performance, a prevalent index named as cross-validation accuracy (CVA) is calculated via a ten-time 3-fold cross-validation.

First, the PRO is tested by varying its parameters for HOME generation, including the maximum scale *S*, the number of directional bins *N*, and the cell grid sizes *P* and *Q*. As seen from the results listed in Table II, it reaches the excellent recognition accuracy that falls between 97% and 99%. Besides, the number of units in each hidden layer is preferably small, which dramatically benefits computational efficiency. Therefore, the PRO reaches a reasonable balance between accuracy and efficiency considering that excessively fine cell grid divisions cause a shortage of grid pixels that constitute meaningful image structures.

Second, the PRO is tested by varying the proportion of samples used for DBN training when respectively parameterized by SETTING_1, SETTING_2, SETTING_3 and SETTING_4, which are four specified settings that result in coarse-grained to fine-grained HOME features and relatively simple to more complicated DBN architectures. From the results listed in Table III, one sees that the minimum accuracy is 92.50% when PRO was trained with one-tenth of the total samples, and the maximum accuracy is 99.82% when it was trained with nine-tenths of the total samples. Moreover, the accuracy remains relatively stable even if the proportion varies from 10% to 90%. This finding shows the stability of PRO in recognition performance.

Third, the PRO is tested in comparison with some typical rival schemes, which are parameterized to attain the maximum accuracy. The results shown in Table IV indicate that the HOME feature outperforms such rival features as the histogram of maximal orientation energy (HMOE) [28], the phase congruency feature (PCF) [21] and the HOG in representation of discriminative cues. Also, the DBN provides better accuracy than such rival classifiers as NN, SVM, RVM and MSRC. This test shows the advantage of the PRO in human recognition over the rivals.

Fourth, the PRO is explored by varying the hidden layer parameters. By manually varying the unit number of the two hidden layers from 3 to 60 and then calculating CVAs of the corresponding PRO, it reaches the curves drawn in Fig. 7. Clearly, PRO produces the best accuracy with its DBN component configured with the MLE estimated number of hidden units. Moreover, by setting the two hidden layer with arbitrary numbers of neuron units and then calculating corresponding CVAs, the best accuracy can be available in Table V. From this table one can only see slightly increasing accuracy. This test demonstrates that the DBN reaches good balance between recognition capability and architectural complexity via construction of the hidden layers with an equal number of neurons, which could be properly estimated by the MLE method.

Fifth, by varying ε from 0.005 to 0.2 and calculating the corresponding CVAs, the influence of the learning rate parameter to the DBN is explored. From the CVA curve drawn in Fig. 8, one sees that the accuracy vibrates slightly round 0.99. It shows the insensitivity of this classifier to this parameter, which assures its classification performance.

Sixth, by varying the epoch from 10 to 200, the CVAs were calculated to explore the influence of training parameter. As observed from the results drawn in Fig. 9, the accuracy roughly keeps stable when the epoch value increases. It demonstrates that the DBN is not sensitive to choices of training parameter.

At last, the PRO is tested on practical infrared sequences. The results from the three sequences introduced in Table I are partially shown in Fig. 10. The rectangle bounding-boxes indicate the correctly recognized blobs, and the eclipse ones denote the falsely ones. As seen, both false positives and false negatives are rare. Such objective indices as true positive, false positive, false negative and precision are calculated and listed in Table VI. The accuracy that exceeds 95% shows excellent adaptation of the PRO to scenic changes.

TABLE I
DETAILS OF THE REPRESENTATIVE TEST SEQUENCES

Sequence No. / Characteristics	Sequence #1	Sequence #2	Sequence #3
Sample frame			
Originality	From OSU Thermal Pedestrian Database	Shot by fixed FLIR A40 camera	Shot by fixed FLIR A40 camera
Noise level	Fair	Low	High
Human sizes	Small	Middle/Large	Middle/Large
Contrast	Low	Good	Low
Boundary	Highly blurred	Blurred	Highly blurred
Inhomogeneity	High	Fair	High

TABLE II
THE CVAS AND HIDDEN UNIT NUMBERS PRODUCED BY THE PRO WITH SPECIFIED HOME GENERATION SETTINGS

Choice of P & Q	P=2,Q=2		P=3,Q=3		P=4,Q=4		P=5,Q=5	
Index value Choice of S & N	CVA	d	CVA	d	CVA	d	CVA	d
S = 1, N=4	0.9412	6	0.9786	7	0.9681	8	0.9677	8
S = 1, N=8	0.9893	9	0.9893	12	0.9946	14	1.0000	14
S = 1, N=12	0.9894	13	0.9731	16	0.9947	18	0.9892	19
S = 1, N=16	0.9893	17	0.9840	20	0.9785	23	0.9946	25
S = 2, N=4	0.9786	6	0.9786	7	0.9892	7	0.9894	8
S = 2, N=8	0.9840	9	0.9839	11	0.9947	13	0.9839	14
S = 2, N=12	0.9946	13	0.9893	14	0.9947	17	0.9893	19
S = 2, N=16	0.9894	17	0.9893	19	0.9839	22	1.0000	23
S = 3, N=4	0.9731	6	1.0000	7	1.0000	7	0.9947	8
S = 3, N=8	0.9947	9	0.9893	11	**0.9977**	12	0.9892	13
S = 3, N=12	0.9893	12	1.0000	14	0.9786	16	0.9893	18
S = 3, N=16	0.9785	15	0.9733	17	0.9786	20	0.9947	21

TABLE III
CVAS OF THE PRO TRAINED BY A VARIABLE PROPORTION OF TOTAL SAMPLES IN THE TEST DATASET

Parameter settings Ratio of training samples	SETTING_1 (S=1,N=4,P=Q=)	SETTING_2 (S=2,N=8,P=Q=3)	SETTING_3 (S=3,N=8,P=Q=4)	SETTING_4 (S=3,N=16,P=Q=5)
10%	0.9250	0.9272	0.9683	0.9905
20%	0.9357	0.9391	0.9908	0.9730
30%	0.9467	0.9454	0.9923	0.9895
40%	0.9259	0.9244	0.9917	0.9908
50%	0.9443	0.9479	0.9989	0.9929
60%	0.9402	0.9420	0.9960	0.9955
70%	0.9411	0.9417	0.9881	0.9887
80%	0.9464	0.9446	0.9741	0.9967
90%	0.9804	0.9821	0.9982	0.9821

TABLE IV
THE CVA COMPARISONS OF PRO WITH THE RIVAL SCHEMES

Scheme	HOG+ NN	HOG+ SVM [10]	HMOE+ NN	HMOE+ SVM [28]	PCF+ RVM[21]
CVA	0.8769	0.8973	0.9120	0.9360	0.9438
Scheme	HOME+ NN	HOME+ SVM	HOME+ RVM	HOME+ MSRC	PRO
CVA	0.9446	0.9660	0.9539	0.9552	0.9845

The experimental results validated PRO in infrared human recognition. The following reasons can be presented.

First, lines/edges have critical roles in providing discriminative cues for human recognition due to the absence of colour and textures in infrared images. Since the MMSR handles well with flat gradients and imaging noise, the HOME feature, which is actually a statistical description of the MMSR, implements good lines/edges representation with the multi-scale conception.

Second, the DBN automatically learns compact features from the HOME, so that it is capable of capturing the multiple layers of abstract features that are more structural and hierarchical by learning the microstructures (such as bright points, dark points and edges) and their distributions as well. This mechanism enables the PRO to reach better understanding of image data and therefore exceeds many shallow models for pattern classification or prediction.

Third, the DBN couples tightly with the HOME feature. For example, the numbers of units of the hidden layers are identically set as the intrinsic dimension of the HOME. In this way, it speeds up the network training because these layer parameters are close to optimal. Moreover, it assures the network convergence by avoiding poorly initialized parameters especially when limited training data are available [29].

TABLE V
THE CVAS OF THE PRO WITH VARYING HIDDEN LAYER PARAMETERS

Unit of Hidden Layers	$D_1=d, D_2=d$	$D_1=d, D_2=2d$	$D_1=d, D_2=3d$	$D_1=d, D_2=4d$
CVA	0.9845	0.9946	0.9946	0.9973
Unit of Hidden Layers	$D_1=2d, D_2=d$	$D_1=2d, D_2=2d$	$D_1=2d, D_2=3d$	$D_1=2d, D_2=4d$
CVA	1.0000	0.9973	0.9919	0.9893
Unit of Hidden Layers	$D_1=3d, D_2=d$	$D_1=3d, D_2=2d$	$D_1=3d, D_2=3d$	$D_1=3d, D_2=4d$
CVA	0.9973	0.9947	0.9973	0.9973
Unit of Hidden Layers	$D_1=4d, D_2=d$	$D_1=4d, D_2=2d$	$D_1=4d, D_2=3d$	$D_1=4d, D_2=4d$
CVA	0.9920	0.9920	1.0000	0.9893

TABLE VI
OBJECTIVE EVALUATION OF PRO IN PRACTICAL HUMAN RECOGNITION.

Index \ Sequence No.	Sequence #1	Sequence #2	Sequence #3
True positives	91	194	318
False positives	4	6	6
False negatives	7	0	29
Precision	0.9579	0.9700	0.9815

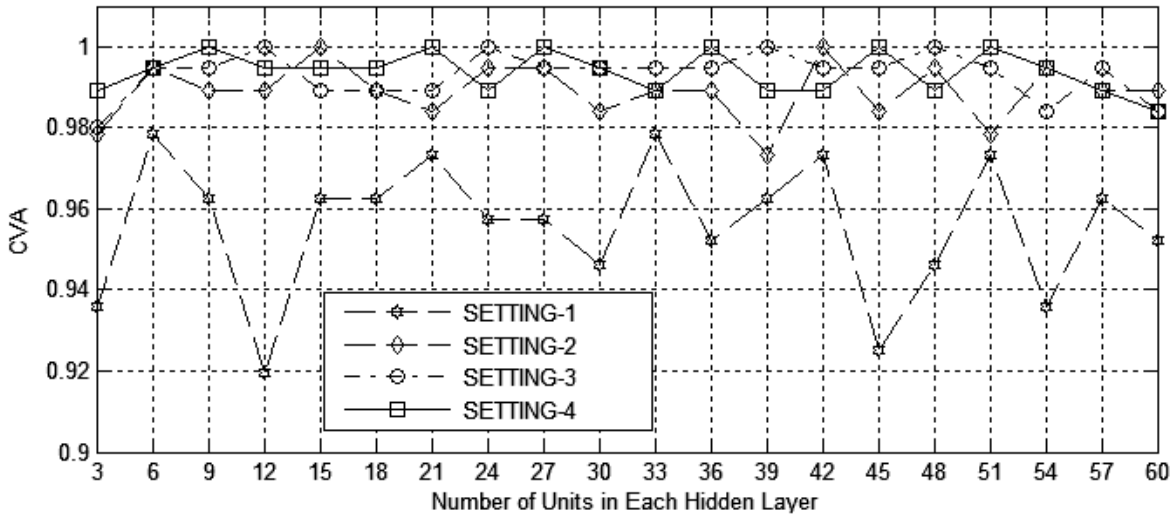


Fig. 7. Comparison of CVAs given by the PRO with specified numbers of units in each hidden layer.

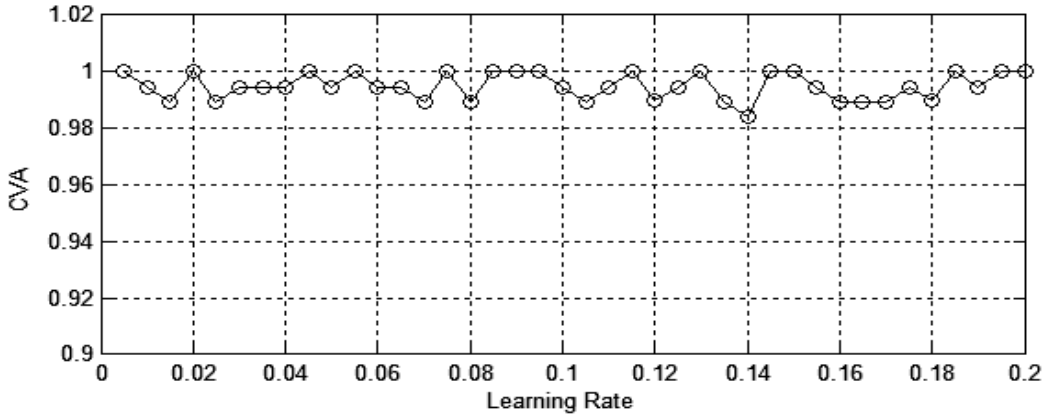


Fig.8. The CVA curve produced by varying the learning rate parameter.

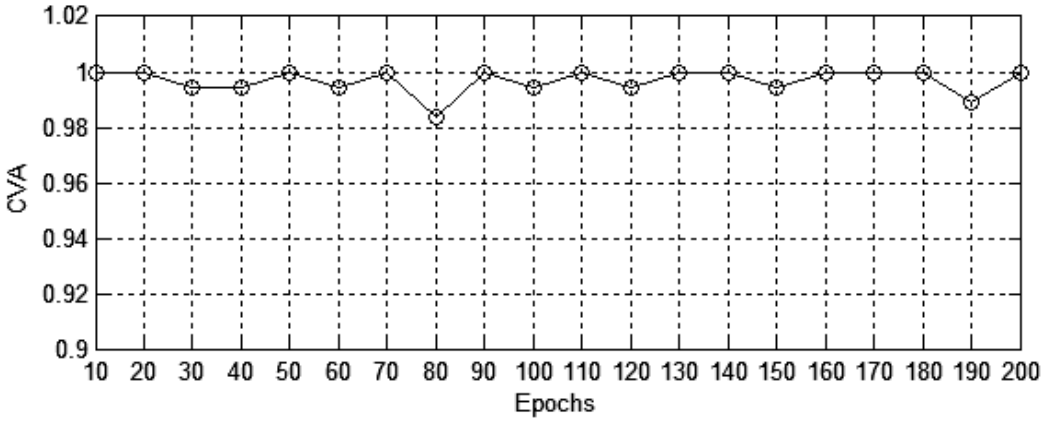


Fig. 9. The CVA curve produced by varying epoch parameter.

(a) Representative results on sequence #1

(b) Representative results on sequence #2

(c) Representative results on sequence #3

Fig. 10. PRO-based recognition results on the test sequences.

VI. Conclusions

A novel method is proposed in this paper for infrared human recognition. The MMSR-based HOME feature provides a satisfactory representation of the discriminative cues on lines/edges in the infrared images. The DBN enables a good understanding of the HOME features. The method has excellent recognition ability and robustness to scenic changes. It is a suitable solution to the infrared human recognition task. In the future, the optimization of algorithmic flow and transfer to real-time platforms should be further explored.

References

[1] Yihui Liang, Han Huang, Zhaoquan Cai, Zhifeng Hao and Kay Chen Tan, "Deep infrared pedestrian classification based on automatic image matting," Applied Soft Computing, vol.77, pp.484-496, 2019.

[2] Jeonghyun Baek, Sungjun Hong, Jisu Kim, Euntai Kim, "Efficient pedestrian detection at nighttime using a thermal camera," Sensors, vol.17, no.8, 1850, 2017.

[3] Hinton, G E, "Deep belief networks," Scholarpedia, vol.4, no.6, pp.804-786, 2009.

[4] Michael Felsberg and Gerald Sommer, "A new extension of linear signal processing for estimating local properties and detecting features," Mustererkennung, pp.195-202, 2000.

[5] Ganggang Dong, Na Wang and Gangyao Kuang, "Sparse representation of monogenic signal: with application to target recognition in SAR images," IEEE Signal Processing Letters, vol.21, no.8, pp.952-956, 2014.

[6] S. Hui, H. Chengying and L. Yupin, "A multi-stage classifier based algorithm of pedestrian detection in night with an infrared camera in a moving car," Proceedings of the Third International Conference on Image and Graphics, pp.120-123, 2004.

[7] Liu T and Liu J, "Mobile robot aided silhouette imaging and robust body pose recognition for elderly-fall detection," International Journal of Advanced Robotic Systems, 2014, vol.11, no.1, 1.

[8] Lamberti F, Santomo R and Sanna A, "Intensity variation function and template matching-based pedestrian tracking in infrared imagery with occlusion detection and recovery," Optical Engineering, vol.54, no.3, 033106, 2015.

[9] Lamberti F, Santomo R, Sanna A and *et al,* "Intensity variation function and template matching-based pedestrian tracking in infrared imagery with occlusion detection and recovery," Optical Engineering, vol.54, no.3, 033106, 2015.

[10] F. Suard, A. Rakotomamonjy, A. Bensrhair, A. Broggi, "Pedestrian detection using infrared images and histograms of oriented gradients," Proceedings of Intelligent Vehicles Symposium, IEEE, pp.206-212, 2006.

[11] Liu Q, Zhuang J and Ma J, "Robust and fast pedestrian detection method for far-infrared automotive driving assistance systems," Infrared Physics & Technology, vol.60, pp.288-299, 2013.

[12] Hua, C., Makihara, Y. and Yagi Y, "Pedestrian detection by using a spatio-temporal histogram of oriented gradients," IEICE Transactions on Information and Systems, vol.96, no.3, pp.1376-1386, 2013.

[13] Wang G and Liu Q, "Far-Infrared based pedestrian detection for driver-assistance systems based on candidate filters, gradient-based feature and multi-frame approval matching," Sensors, vol.15, no.12, pp.32188-32212, 2015.

[14] Zhao X, He Z, Zhang S and *et al*, "Robust pedestrian detection in thermal infrared imagery using a shape distribution histogram feature and modified sparse representation classification," Pattern Recognition, vol.48, no.6, pp. 1947-1960, 2015.

[15] Dong Xia, Hao Sun and Zhenkang Shen, "Real-time infrared pedestrian detection based on multi-block LBP," 2010 International Conference on Computer Application and System Modeling (ICCASM 2010), Taiyuan, pp. V12-139-V12-142, 2010.

[16] Ko B , Kim D and Nam J, "Detecting humans using luminance saliency in thermal images," Optics letters, vol.37, no.20, pp.4350-4352,2012.

[17] Besbes B, Rogozan A and Rus A M, "Pedestrian detection in far-infrared daytime images using a hierarchical codebook of SURF," Sensors, vol.15, no.4, pp.8570-8594, 2015.

[18] Liu Feng,Wang Sibo,Wang Xiangjun,Zhao Guangwei and Huo Wenjia, "Infrared pedestrian detection method in low visibility environment based on multi feature association," Infrared and Laser Engineering, vol.47, no.6, 0604001, 2018.

[19] Lahmyed R , El Ansari M , Ellahyani A, "A new thermal infrared and visible spectrum images-based pedestrian detection system," Multimedia Tools and Applications, vol.78, no.12, pp.15861-15885, 2019.

[20] Diulhio D O , Marco W, "Using deep learning and low-cost RGB and thermal cameras to detect pedestrians in aerial images captured by multirotor UAV," Sensors, vol.18, no.7, 2244, 2018.

[21] Tan Y, "The phase congruency based feature for thermal infrared human detection," ICIC Express Letters, vol.9, no.6, pp.1747-1752.

[22] Zhang X, Chen G, Saruta K and *et al*, "Deep convolutional neural networks for all-day pedestrian detection," International Conference on Information Science and Applications. Springer, Singapore, pp.171-178, 2017.

[23] Hinton G E, "Training products of experts by minimizing contrastive divergence," Neural Computation, vol.14, no.8, pp.1771-1800, 2002.

[24] Jie Li, Mingqiang Yang, Yupeng Liu, Yanyan Wang, Qinghe Zheng, and Deqiang Wang, "Dynamic Hand Gesture Recognition Using Multi-direction 3D Convolutional Neural Networks," Engineering Letters, vol. 27, no.3, pp490-500, 2019.

[25] Levina E, Bickel P J. Maximum likelihood estimation of intrinsic dimension [J]. Advances in Neural Information Processing Systems, 2004, 17:777-784.

[26] HINTON G E, OSINDERO S, THE Y-W. A fast learning algorithm for deep belief nets [J]. Neural Computation, 2006, 18(7): 1527-1554.

[27] Davis J W and Keck M A, "A two-stage template approach to person detection in thermal imagery," Application of Computer Vision, 2005. WACV/MOTIONS '05 Volume 1. Seventh IEEE Workshops on. IEEE, 2005.

[28] Yun T J, Guo Y C and Chao G, "Human detection in far-infrared images based on histograms of maximal oriented energy map," International Conference on Wavelet Analysis and Pattern Recognition. IEEE, pp.933-938, 2008.

[29] Larochelle H, Erhan D, Courville A and *et al*, "An empirical evaluation of deep architectures on problems with many factors of variation," International Conference on Machine Learning. ACM, pp.473-480, 2007.

Yong Tan was born in 1981, in Sichuan province, China. He received his M.S. degree in information & communication engineering and Ph.D. degree in instrumental science & technology from Chongqing University, Chongqing, China, in 2007 and 2013, respectively. His research interests include image processing and pattern recognition.

He is now an associate professor at Yangtze Normal University, Chongqing, China.