A Motion Deviation Image-based Phase Feature for Recognition of Thermal Infrared Human Activities

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

***Abstract*—A novel activity feature for thermal infrared human activity recognition, which suffers from poor quality infrared imaging and a great variation in human subjects, is proposed in this paper. Extraction of this feature consists of three major stages. First, a coarse-to-fine localization procedure extracts the regions of interest (ROIs) that may contain human subjects from a raw infrared sequence. Second, it generates a motion deviation image (MDI), which is a novel spatio-temporal template, from the available ROIs to represent the infrared activity sequence efficiently. Third, from such MDI, it generates a directional phase congruency-based feature that codes the significant activity information via the steps including calculating phase maps, estimating intrinsic dimensions and reducing vector dimensionality. The proposed feature is validated on the IADB infrared human activity database, and the experimental results show its advantages in recognition accuracy when used for multi-class activity recognition due to its robustness to poor quality thermal infrared imagery, good representation of activity information and effective removal of noise.**

***Index Terms*—thermal infrared human activity recognition, motion deviation image, directional phase congruency, intrinsic dimension estimation, reduction in dimensionality**

I. Introduction

Human activity recognition (HAR) has attracted much attention due to the number of applications for public security, traffic monitoring, and military actions. Regularly, HAR systems are configured with imaging sensors that work in the visual light spectrum, as these sensors are capable of providing good quality images of enriched information. However, HAR systems might degrade or even fail when these sensors suffer from such troubles as shadows and poor lighting. Instead, thermal infrared imaging sensors, which detect objects by their thermal radiations, deal well with bad light conditions and can thereby make HAR systems work constantly. However, developing thermal infrared HAR systems faces two major challenges. The first challenge is the poor quality of infrared imaging, including low-contrast, boundary-blurring, loss of colour and low resolution, and the second is the considerable variations in human subjects on aspects of pose, size and motion patterns. These challenges impose considerable difficulties in reaching acceptable HAR accuracy.

General HAR involves extracting activity features and classifying activities [1]. In the feature extraction step, a reasonable and meaningful descriptor is formulated from raw data that is regularly in the form of an image or image sequence. This descriptor identifies the intrinsic differences between different activities or different observations of an identical activity. In the activity classification step, activity observations are labelled by supervised or non-supervised classifiers. The classifiers are expected to have good generalization to adapt to new observations. Although both steps have a dramatic influence on HAR performance, the availability of effective activity features is generally viewed as fundamental and critical.

A novel feature for HAR via thermal infrared imagery is proposed in this paper. Through a procedure that consists of localization of activity subjects, generalization of a spatio-temporal template named the motion deviation image (MDI) and feature extraction from MDI, it identifies as much activity information as possible from an infrared activity sequence characterized by poor imaging quality and complex patterns of human activities, and therefore, it demonstrates advantages in recognition performance.

In the rest of this paper, the related literature is briefly covered in section II. In section III, the formulation of the proposed feature is given in detail. Experimental results and algorithmic analysis are provided in section IV, and some concluding remarks are presented in section V.

II. Related Work

Many activity features have already been proposed for HAR, and they can be roughly divided into body modelling, local representation, global representation and multi-feature fusion categories. Body modelling features [2] [3] [4] formulate two-dimensional or three-dimensional models, such as the rectangle and skeleton models, to describe human bodies. Due to the large degree of freedom in a human body, they suffer from bad generality and the weakness to mutual occlusions. Local representation features generally handle occlusions and crowd scenarios well, as they provide a holistic representation of an activity by incorporating a large number of local information units. The units, each of which provides partial and weak activity information, can be of such

Manuscript received May 13, 2019; revised July 29, 2019. 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.

forms as key-point [5][6][7][8], trajectory [9][10][11], orient gradient [12][13][14], and bag-of-words [15][16][17]. Excessive time consumption is often required for calculating this kind of features. By contrast, global representation features do not concern local units but the whole appearance or motion of an actor, and generating these features can be much more time-saving. Representative features of this kind include human silhouette [18] [19] and energy template. In particular, the energy template, including the variation energy image (VEI) [1], the motion history image (MHI) [20], the silhouette energy image (SEI) [21] and the gait energy image (GEI) [22], provides efficient representation of spatio-temporal information of an activity with very reasonable calculation cost. The fusion-based features [23] [24] [25] consist of multiple features originating from different sensors, scales, viewpoints, and decisions. Because they overcome the limitation of any single feature, they generally promote recognition precision if the multiple sources are truly complementary. The major challenges lie in the choice of fusion techniques and the increasing calculation cost. Generally, compared with the rival features, global representation features perform better in balancing recognition accuracy and computational cost. From this viewpoint, a new energy template, i.e., the MDI, is proposed in this paper, and then the phase-based feature is formulated from this template for thermal infrared HAR.

III. The Proposed Feature Descriptor

Fig 1 presents the generation flow of the proposed feature descriptor. It can be divided into three major stages. In this first stage, the regions of interest (ROIs) that may contain the activity subjects are searched from a thermal infrared activity sequence. In the second stage, a MDI, which provides a compressed representation of the activity sequences, is constructed for incoming extraction of activity information. In the last stage, a phase-based feature that represents intrinsic activity information is extracted from the lines/edges of the MDI and it is readily used for incoming activity recognition.

A. Stage #1: Localization of Activity Subjects

Localization of activity subjects means determining the sub-regions in which the human subjects of interest appear in activity sequences. It is a fundamental but difficult task prior to activity recognition, due to the poor quality of infrared imaging and the inhomogeneity of human subjects in intensity. To avoid fragmentation, a coarse-to-fine localization method, which consists of operations including calculating the difference motion history image (DMHI) [21], one-dimensional searching localization, and size normalization, is proposed.

The method starts to construct a DMHI. Focusing on the history information in the image regions in motion, the DMHI is advantageous to handle complicated activities and incomplete subject silhouettes well. Originally, the DMHI was defined as the sum of the absolute difference between adjacent frames in the temporal duration t. By fixing the parameter t as the period of an activity and replacing it by N, which is the frame number of the clip that covers the period, the DMHI can be calculated as follows

$$DMHI(x,y)=\sum_{n=1}^{N-1}|\,f_n(x,y)-f_{n+1}(x,y)\,| \qquad (1)$$

where $f_n(x,y)$ denotes the nth frame of the activity clip that covers the very period. Fig 2 shows two exemplar DMHIs for the activity "running" and "kicking", and it sees clearly that the subjects appear in the saliency regions.

Over the DMHI, one-dimensional searching runs to coarsely locate the activity subjects. In its horizontal-first search, a pixel-vertical-projection curve is firstly calculated by summarizing the grey-level intensities of DMHI pixels along the vertical direction. Then, in the DMHI the vertical strips that correspond to the peaks of this one-dimensional curve are searched. In its vertical-second search, a pixel-horizontal-projection curve is calculated, and then in the DMHI, the horizontal strips that correspond to the peaks of the new one-dimensional curve are also searched. Now, the activity subjects can be localized in the regions intersected by the available horizontal strips and vertical strips. Note that the parameter that determines the peaks of each one-dimensional curve can be chosen as the average of the elements of the very curve. Next, to locate the activity subjects that appear in the frame images, the horizontal one-dimensional search runs repeatedly over each frame image in the identical subdomain with that of the searched intersection regions of the DMHI. In this way, the subjects of interest get localized quickly and accurately. To eliminate the effect of variation in imaging distances, the subject sizes are normalized at last. An illustration of this coarse-to-fine localization method has been shown in Fig 3.

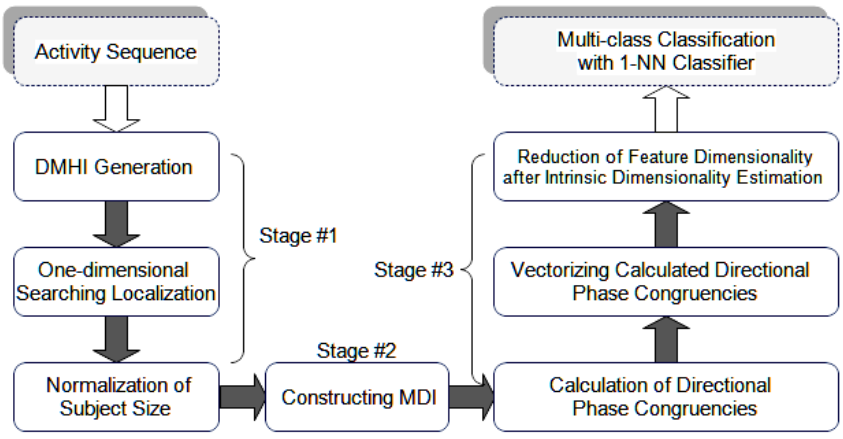
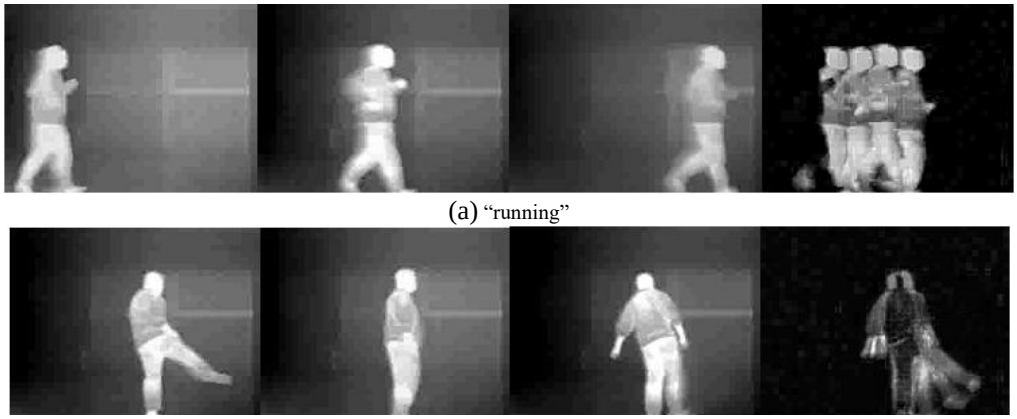


Fig 1. Flow of generating the proposed activity feature.



(a) "running"

(b) "kicking"

Fig 2. Exemplar activity frames and the DMHIs for (a) "running" and (b) "kicking".

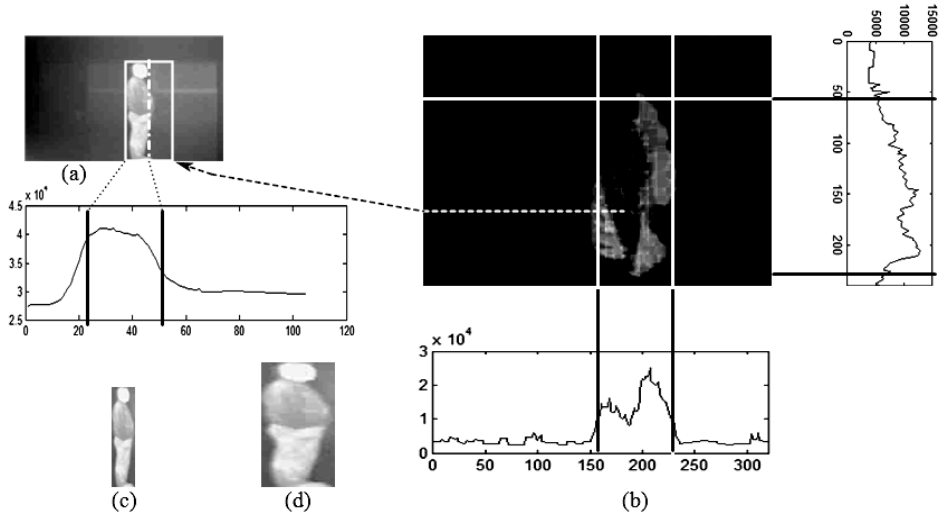
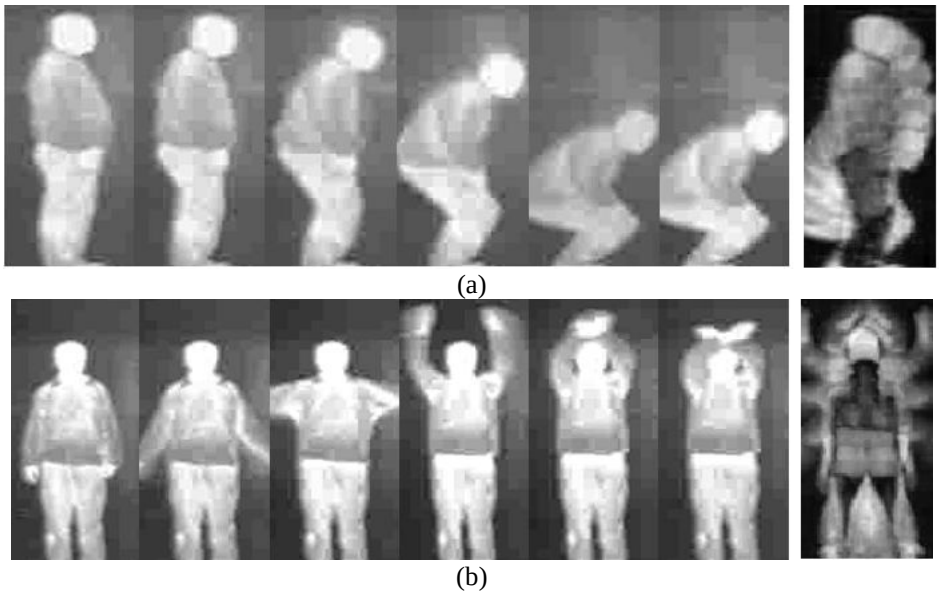


Fig 3. Exemplar images illustrating the one-dimensional searching localization method. (a) One sample frame of the activity "bending". (b) The DMHI for "bending" and its grey-level projection curves. (c) Localized activity subject appearing in the frame image (a). (d) Size-normalized activity subject.



(a)

(b)

Fig 4. Exemplar ROIs for (a) "bending" (b) "waving" and their MDIs.

B. Stage #2: Generalization of a New Spatio-Temporal Template: MDI

Let $B_n(x, y)$ denote the subject region of interest (ROI) extracted from the n th frame. For an activity clip having N frames that lasts a period, the average of these ROIs is

$$U(x,y) = \frac{1}{N}\sum_{n=n_0}^{n_0+N-1} B_n(x,y) \tag{2}$$

and their deviation is

$$\sigma(x,y) = \sqrt{\frac{1}{N-1}\sum_{n=n_0}^{n_0+N-1}[B_n(x,y)-U(x,y)]^2} \tag{3}$$

Now, the template MDI can be defined as the normalized $\sigma(x, y)$, namely,

$$MDI(x, y) = \frac{\sigma(x, y)}{\sigma_{\max}} \quad (4)$$

where $\sigma_{\max}$ is the maximum element of the matrix σ. In Fig 4, it shows two MDIs related to "bending" and "waving". Clearly, they depict activities in a highly compressed way.

C. Stage #3: Feature Extraction from MDI.

As shown in Fig. 4, a MDI epitomizes the spatio-temporal information of an activity. Since its edges primarily represent the shape profile and the motion characteristics of an activity, it is reasonable to extract activity features from such edges for activity recognition. In this subsection, phase congruency [26], which is powerful for line and edge description, is utilized for this purpose and it leads to a phase-based activity feature descriptor.

Under the conception of multi-resolution, phase congruency is defined as follows:

$$\begin{cases} PC(x,y) = \dfrac{\sum_o \sum_s W_o \lfloor A_{so}(x,y)\Delta\phi_{so}(x,y) - T_o(x,y) \rfloor}{\sum_o \sum_s A_{so} + \varepsilon} \\ A_{so}(x,y) = \sqrt{(I(x,y) * M_{so}^{even}(x,y))^2 + (I(x,y) * M_{so}^{odd}(x,y))^2} \\ \phi_{so}(x,y) = \arctan((I(x,y) * M_{so}^{even}(x,y)) / (I(x,y) * M_{so}^{odd}(x,y))) \\ \Delta\phi_{so}(x,y) = \cos(\phi_{so}(x,y) - \bar{\phi}_o(x,y)) - |\sin(\phi_{so}(x,y) - \bar{\phi}_o(x,y))| \end{cases} \quad (5)$$

where $I(x, y)$ represents the image over which the phase calculation is performed, M_{so}^{even} and M_{so}^{odd} are the respective even-symmetric and odd-symmetric filters at the scale s, $s = 1, 2, ..., p$ and the orientation o, and $o = 1, 2, ..., q$. The symbol "*" is a convolution operator, W_o represents the weights for the frequency spread, T_o is the estimated noise energy at orientation o, ε is a small constant that prevents division by zero. Additionally, $\lfloor \bullet \rfloor$ is an operator that enables the enclosed quantity equal to itself if the value is positive and zero if not, $\Delta\phi_{so}$ is the phase deviation measure, and $\bar{\phi}_o$ is the mean phase angle at the orientation o.

Based on above definition of the phase congruency, the concept of directional phase congruency can be defined by

$$PC_o(x,y) = \frac{\sum_s W_o \lfloor A_{so}(x,y)\Delta\phi_{so}(x,y) - T_o(x,y) \rfloor}{\sum_s A_{so} + \varepsilon}, \; o = 1, 2, ..., q. \quad (6)$$

In (6), the even-symmetric and odd-symmetric filters, i.e., M_{so}^{even} and M_{so}^{odd}, are chosen as the two-dimensional log-Gabor filter banks $\{G_{so} : s = 1, 2, ..., p; o = 1, 2, ..., q\}$, considering their adaptation to the human visual system.

The directional phase congruency provides phase representation of image edges along multiple orientations. It is reasonably adopted to describe MDI edges in order to reach a fine representation of the variations in human shapes and activity motions.

By calculating PC_o, $o = 1, 2, ..., q$ and dividing each of the PC_o maps into $M \times N$ non-overlapping cells, a series of cell averages can be calculated. Then, irregularly, these averages can be sequentially collected by

$$H_o = \{m_1, m_{2,} ... m_{M \times N}\}, o = 1, 2, ..., q. \quad (7)$$

Moreover, by pixel-by-pixel comparison of the q directional phase congruency maps, a maximum map MAX_PC can be calculated as follows

$$MAX\_PC(x,y) = \max(PC_1(x,y), PC_2(x,y), ..., PC_q(x,y)) \quad (8)$$

and it leads to the cell average vector

$$MAX\_H = \{n_1, n_{2,} ... n_{M \times N}\}. \quad (9)$$

Finally, collecting H_o, $o = 1, 2, ..., q$ and MAX_H, it produces

$$H_L = \{H_1, H_2, ..., H_q, MAX\_H\} \quad (10)$$

Note, H_L may be of high dimensionality due to the parameters related to the log-Gabor filter banks and cell division grids. Moreover, it may be polluted by spurious information and noise originating from poor imaging quality and MDI generation errors as well. To overcome these problems, dimensionality reduction can be applied. However, before it runs, the maximum likelihood estimator (MLE) [27] is used to estimate the intrinsic dimensionality of H_L considering that it provides sufficient accuracy with reasonable computational cost. With the estimated dimensionality, the technique linear discriminant analysis (LDA) [28] runs to map H_L to a new feature descriptor H that is of the estimated dimensionality. This descriptor H, which describes H_L without loss of significant information, is ready for activity recognition and would lead to better class-separability via avoidance of overfitting and lower computational cost. Taking the MDI for "waving" as an example, Fig 5 illustrates the generation of the descriptor H.

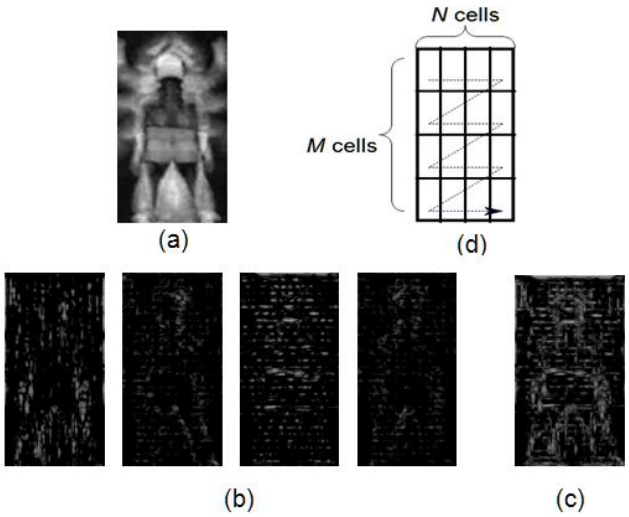


Fig 5. Exemplar images to illustrate feature extraction from an MDI. (a) The MDI for "waving". (b) Four directional phase maps. (c) Maximum phase map. (d) Cell division and zigzagged collection of averages.

IV. Experiment Results and Algorithmic Analysis

The proposed feature (PRO) was implemented by MATLAB 2013b on a personal computer with Intel Core i7-4720HQ 2.60 GHz CPU, 4G RAM and Windows 10 operating system. To test its performance, 12 different activities, each of which was performed by 10 different human subjects, were selected from the IADB infrared human activity dataset [21]. The activities were bending, kicking, punching, jumping-jack, jumping forward on two legs, jumping in place on two legs, running, galloping sideways, skipping, walking, one-hand waving and two-hands waving. Some sample frames of these activities are provided in Fig 6. Parameterized by $p = 1$, $q = 4$, and

$M = N = 4$, and configured with the nearest neighbour classifier (NN), which is viewed as the most light-weighted classifier, PRO was validated via a ten-time 3-fold cross-validation process.

To objectively evaluate the multi-class discrimination performance, confusion matrix was first calculated for PRO. In this matrix, the data located at the main diagonal cells of the matrix represent the percentage of correct recognition and that the intersection of the i'th row and the j'th column represents the percentage of class i activities being recognized as class j. In other words, the diagonal data represent the recognition accuracy of different activities, while others are the percentages of misclassification. As seen from the calculated matrix given in Table 1, the only confusion occurs between "p-jump" and "kick" due to their similarity, while most activities receive perfect recognition.

Table 1. Confusion matrix with PRO.

	bend	jack	jump	kick	pjump	punch	run	side	skip	walk	wave1	wave2
bend	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
jack	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
jump	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
kick	0.00	0.00	0.00	0.90	0.10	0.00	0.00	0.00	0.00	0.00	0.00	0.00
pjump	0.00	0.00	0.00	0.20	0.80	0.00	0.00	0.00	0.00	0.00	0.00	0.00
punch	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
run	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
side	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00
skip	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
walk	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00
wave1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
wave2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

Next, PRO was tested in comparison with such powerful rival feature descriptors as GEI [22], 3D gradient [29], SEI-HOG [21] and VEI-LogGabor [1], and the index named as cross-validation accuracy (CVA) [29] was adopted as the evaluation index. From the results given in Table 2, it sees that PRO achieves an accuracy of 98.3% and outperforms the rivals by at least 4.0%. Moreover, the index is calculated for the PRO configured with such representative classifiers as a back-propagation neural network (BPNet) [30], support vector machine (SVM) [31], random forest and AdaBoost. In this test, preferable parameters were chosen for the classifiers to reach as high accuracy as possible. The results given in Table 3 show that PRO achieves the accuracy that are higher than 96% when arbitrarily working with one of these classifiers differencing in working principle and classification ability. It demonstrates that PRO achieves good representations of intrinsic activity cues and is resultantly insensitive to choices of classifiers.

To examine the effect of parameter settings on PRO, this feature descriptor is generated with different choices of scale number p, orientation number q and grid size, and then applied for recognition. The corresponding evaluation results have been given in Table 4. From this table, the following phenomena can be seen. First, the sizes of cell grids obviously affect the recognition accuracy, and the choices of relatively small sizes would be preferable. Second, a relatively larger q results in better accuracy if cell grid sizes are fixed. It means that more precise calculation of the directional phase makes better representation of the MDI edges on which the activity cue concentrates. Third, along with increasing p, the recognition accuracy ascends but starts to descend if p meets a certain turn-point value. It shows that superfine scale settings do not absolutely provide better edge representation, due to the noise originating from infrared imagining and MDI generation. In Table 5, the dimensions of the parameterized PRO descriptors are given. Obviously, the relatively small dimension values are definitely beneficial to the efficiency of incoming activity classification. To verify the values, the feature vectors with manually changed dimensions are tested and then it reaches the CVA curves shown in Fig 7. An arbitrary curve in this figure shows that the recognition accuracy ascends rapidly to an approximate maximum and then roughly keeps such maximum when the dimension of the very feature vector varies from 1 to its allowable maximum dimension. As the turn-point dimension that corresponds to the foremost maximum accuracy accord with the ones given in Table 5, the estimation of intrinsic dimensionality by MLE is accurate.

Generating the dataset of PRO under the situation with arbitrary one of the three parameter settings, which are respectively named as SETTING_1, SETTING_2 and SETTING_3, and then training NN using randomly selected samples that take up 10% to 90% of the total samples in the dataset, the average CVAs can be calculated by repeatedly applying this classifier to recognize the left samples in the dataset. From Table 6, in which the results are given, it sees the following phenomena. First, even if only 10% of the samples are used for training the classifier, the recognition accuracy can still be satisfactory. Second, although the ratio of training samples varies drastically from 10% to 90%, the recognition accuracy only slightly changes regardless of the parameter settings that represent coarse-grained to fine-grained PRO. The phenomena demonstrate that PRO has strong discriminative capability and therefore enables a classifier to be well trained by just using a few samples.

Finally, the average time costs of recognizing one sample are recorded for several PRO-based recognition schemes, and the results, which are listed in Table 7, show great discrepancy of these schemes in efficiency. With reference to this table and previous Table 3, it sees that PRO does benefit HAR performance in efficiency, as it enables lightweight classifiers, which generally run faster, good enough in providing satisfactory accuracy. Therefore, such lightweight classifiers as the NN are preferable for PRO based HAR systems.

The experimental results above validate the advantages of PRO for the following reasons.

First, in comparison with such rival templates as SEI and VEI, the MDI has better performance in faithfully conserving spatial-temporal information of an activity because of two strategies. The first strategy is that the ROIs used to formulate this template are generated by a projection-based method instead of ordinary segmentation methods. In this way, it avoids loss of information caused by undesirable image segmentation. The second strategy is that the MDI is directly generated from a series of same-sized grey-level ROIs instead of binary ROIs. Resultantly, it preserves the information concerning subject appearances and motion variations as well.

Second, because the responses of the 2D log-Gabor filter bank over multiple image scales and orientations are maximal in phase at significance points, phase congruency provides good representations of MDI lines/edges. In comparison with gradient-based edge detectors, the phase-based descriptor

demonstrates stronger robustness to image variations originating from illumination and blurring due to its multi-resolution calculation approach and the detection of lines (or edges) by phase rather than magnitude.

Third, with the accurately estimated intrinsic dimension, the LDA helps to remove as much noise originating from thermal imagining and MDI formulation as possible. Additionally, it avoids possible singularity problems in activity classification when the number of training samples per class is significantly smaller than that of sample elements.

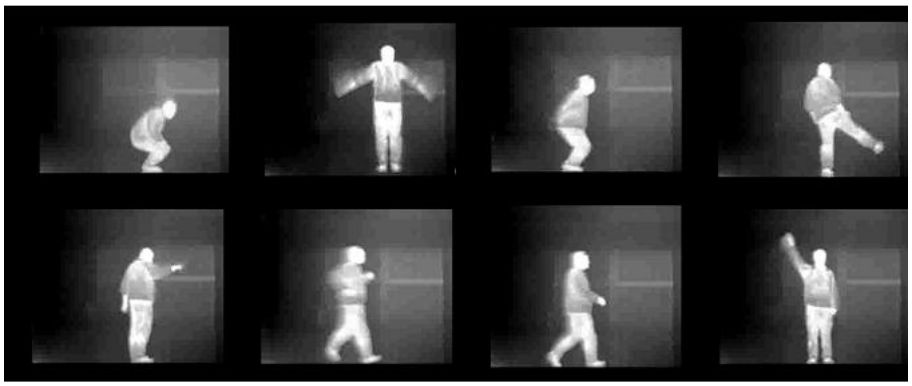


Fig 6. Sample frames of the activities in the IADB database

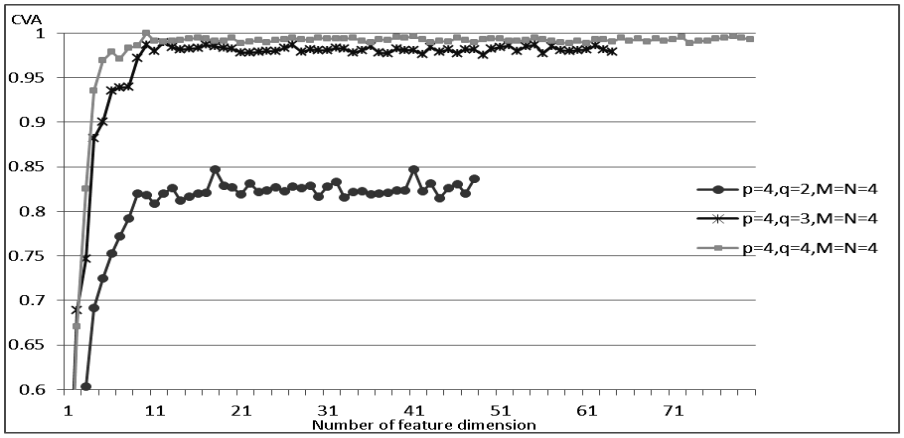


Fig 7. CVA curves of the feature vectors featured by variable dimension.

Table 2. CVA comparisons of PRO with four rival features.

Method	GEI+NN[22]	3D gradient+NN[29]	SEI-HOG+NN[21]	VEI-LogGabor+RVM [1]	PRO+NN
CVA	0.810	0.832	0.849	0.944	0.983

Table 3. CVA comparisons of the recognition schemes composed of PROand five representative classifiers

Method	PRO+NN	PRO+SVM	PRO+BPNet	PRO+Random Forest	PRO+AdaBoost
CVA	0.983	0.964	0.971	0.991	0.993

Table 4. CVAs of differently parameterized PRO.

$M \times N$	2×2				3×3				4×4			
p \ q	2	3	4	5	2	3	4	5	2	3	4	5
1	0.148	0.522	0.508	0.587	0.693	0.829	0.921	0.978	0.796	0.948	0.983	0.985
2	0.473	0.531	0.600	0.665	0.781	0.788	0.962	0.959	0.874	0.968	0.985	0.993
3	0.358	0.650	0.705	0.731	0.816	0.920	0.949	0.981	0.936	0.977	0.985	0.991
4	0.339	0.567	0.649	0.383	0.659	0.917	0.744	0.955	0.949	0.981	0.988	0.996
5	0.523	0.671	0.683	0.714	0.782	0.639	0.946	0.982	0.868	0.956	0.976	0.986

Table 5. The dimensionalities with PRO under specified parameter settings

$M \times N$	2×2				3×3				4×4			
p \ q	2	3	4	5	2	3	4	5	2	3	4	5
1	5	6	6	7	7	7	8	9	8	8	9	9
2	5	6	7	7	7	8	9	9	8	9	10	10
3	5	6	7	8	8	8	9	10	9	10	11	11
4	5	6	7	8	8	9	10	10	10	11	13	12
5	5	6	8	8	8	8	10	11	10	11	12	13

Table 6. The average CVAs of recognition schemes composed of the PRO under three parameter settings and the NN classifier trained by variable proportions of samples in the test dataset

Parameter settings / Ratio of training samples	SETTING_1 (p=1,q=2,M=N=4)	SETTING_2 (p=1,q=4,M=N=4)	SETTING_3 (p=2,q=2,M=N=6)
10%	0.758	0.958	0.991
20%	0.766	0.975	0.992
30%	0.775	0.977	0.994
40%	0.789	0.981	0.995
50%	0.790	0.985	0.996
60%	0.793	0.986	0.998
70%	0.781	0.988	0.997
80%	0.752	0.989	0.998
90%	0.739	0.990	0.998

Table 7. Efficiency Comparisons of the PRO-based recognition schemes.

Method	PRO+NN	PRO+SVM	PRO+BPNet	PRO+Random Forest	PRO+AdaBoost
Time Cost(/second)	3.75e-5	5.42e-5	6.2e-5	5.10e-4	1.85e-3

V. Conclusions

In this paper, an MDI-based phase feature descriptor is presented for infrared HAR. Because of the techniques including the spatial-temporal template MDI, the directional phase congruency-based description of activity information, and MLE and LDA based removal of noise, the feature descriptor provides a good representation of human activity cues. When parameterized properly, it leads to dramatically better multi-class recognition accuracy than rivals. Moreover, as lightweight classifiers work well with the feature, it facilitates the formulation of fast HAR schemes. These advantages make it preferable for HAR systems. In the future, algorithmic optimization and transfer to specified hardware systems should be explored.

References

[1] He W H, Yow K C, and Guo Y C, "Recognition of human activities using a multiclass relevance vector machine," Optical Engineering, vol. 51, no. 1, 017202, 2012.

[2] Ikizler N and Duygulu P, "Histogram of oriented rectangles: A new pose descriptor for human action recognition," Image & Vision Computing, vol. 27, no. 10, pp1515-1526, 2009.

[3] Raza M., Chen Z., Rehman S. U., Peng W., and Peng B., "Appearance based pedestrians' head pose and body orientation estimation using deep learning," Neurocomputing," Neurocomputing, vol. 272, pp 647-659, 2018.

[4] Wu W, Yang Y, Liu R, and Deng C., "Joint-based multi-task sparse learning for human action recognition," International Conference on Internet Multimedia Computing & Service. ACM, pp1-4, 2015.

[5] Laptev, I, "On space-time interest points," International Journal of Computer Vision, vol. 64, no. 2-3, pp107-123, 2005.

[6] APA Chakraborty B., Holte M. B. , Moeslund T. B., and Jordi Gonzàlez, "Selective spatio-temporal interest points," Computer Vision and Image Understanding, vol. 116, no. 3, pp 396-410, 2012.

[7] Scovanner P.,Ali S., and Shah M, "A 3-dimensional sift descriptor and its application to action recognition," in proceedings of the Fifteenth ACM International Conference on Multimedia, pp357-360, 2007.

[8] Mouna S, Mounim E Y, and Bernadette D, "A two-layer discriminative model for human activity recognition," IET Computer Vision, vol. 10, no. 4, pp273-278, 2017.

[9] Shao Y H, Guo Y C and Gao C, "Infrared human action recognition using dense trajectories-based feature," Journal of Optoelectronics Laser, 2015, vol. 26, no. 4, pp. 758-763.

[10] Guo Y, Li Y F, and Shao Z P, "DSRF: A flexible trajectory descriptor for articulated human action recognition," Pattern Recognition, vol. 76, no. 4, pp137-148, 2018.

[11] Guan Y P and Mao W Q, "Pedestrian virtual space based abnormal behavior detection," IAENG International Journal of Computer Science, vol. 46, no. 2, pp311-320, 2019.

[12] Chen C, Zhang B, Hou Z, Jiang J, and Yang Y, "Action recognition from depth sequences using weighted fusion of 2D and 3D auto-correlation of gradients features," Multimedia Tools and Applications, vol. 76, no. 3, pp 4651-4669, 2016.

[13] Wang, H.,Klaeser, A.,Schmid, C., and Liu C L, "Dense trajectories and motion boundary descriptors for action recognition," International Journal of Computer Vision, vol. 103, no. 1, pp. 60-79, 2013.

[14] Li N, Cheng X, Zhang S, and Wu Z, "Realistic human action recognition by Fast HOG3D and self-organization feature map," Machine Vision and Applications, vol. 25, no. 7, pp1793-1812, 2014.

[15] Iosifidis A.,Tefas A and Pitas I, "Discriminant Bag of words based representation for human action recognition," Pattern Recognition Letters, vol. 49, pp185-192, 2014.

[16] Somasundaram G, Cherian A, Morellas V, and Papanikolopoulos N, "Action recognition using global spatio-temporal features derived from sparse representations," Computer Vision and Image Understanding, Jun, vol.123, pp1-13, 2014.

[17] Wang Y, Shi Y, and Wei G, "A novel local feature descriptor based on energy information for human activity recognition," Neurocomputing, vol. 228, no. 3, pp19-28, 2017.

[18] Qian H, Zhou J, Mao Y, and Yuan Y, "Recognizing human actions from silhouettes described with weighted distance metric and kinematics," Multimedia Tools and Applications, vol. 76, no. 21, pp 21889-21910, 2017.

[19] Kushitwaha A K S, Srivastava S, and Srivastava R, "Multi-view human activity recognition based on silhouette and uniform rotation invariant local binary patterns," Multimedia Systems, vol. 23, no. 4, pp451-467, 2017.

[20] Ahad M A R, Tan J K, Kim H, and Ishikawa S, "Motion history image: its variants and applications," Machine Vision and Applications, vol. 23, no. 2, pp 255-281, 2012.

[21] Li J F and Gong W G, "Application of thermal infrared imagery in human action recognition," Nanotechnology and Computer Engineering, no. 121-122, pp368-372, 2010.

[22] Han J and Bhanu B, "Individual recognition using gait energy image," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 28, no. 2, pp316-322,2006.

[23] Leite D Q, Duarte J C, Neves L P, Oliveira J C, and Giraldi G A, "Hand gesture recognition from depth and infrared Kinect data for CAVE applications interaction," Multimedia Tools and Applications, vol. 76, no. 20, pp20423-20455, 2017.

[24] Qi W, Han J, Zhang Y, and Bai L F, "Infrared object detection using global and local cues based on LARK," Infrared Physics & Technology, vol. 76, pp206-216, 2016.

[25] Al-Temeemy and Ali A, "Multispectral imaging: Monitoring vulnerable people," Optik - International Journal for Light & Electron Optics, vol. 180, no. 15, pp469-483, 2019.

[26] Kovesi P, " Image features from phase congruency," Videre: Journal of Computer Vision Research, vol. 1, no. 3, pp1-26, 1999.

[27] Bouveyron C, Celeux G, and Stéphane Girard, "Intrinsic dimension estimation by maximum likelihood in isotropic probabilistic PCA" Pattern Recognition Letters, vol. 32, no. 14, pp1706-1713, 2011.

[28] Jian Y. Theory of Fisher Linear Discriminant Analysis and Its Application [J]. Acta Automatica Sinica, vol.29, no.4, pp481-493, 2003.
[29] Dollar P, Rabaud V, Cottrell G, and Belongie S, "Behavior recognition via sparse spatio-temporal features," Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, pp65-72, 2005.
[30] R.P.W. Duin, P. Juszczak, P. Paclik, E. Pekalska, D. de Ridder, D.M.J. Tax and S. Verzakov, "PRTools4.1, A Matlab Toolbox for Pattern Recognition," Delft University of Technology, 2007.
[31] Chih-Chung Chang and Chih-Jen Lin, "LIBSVM : a library for support vector machines," ACM Transactions on Intelligent Systems and Technology, vol.2, no.3, pp27:1-27:27, 2011.

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 a full-time lecturer at Yangtze Normal University, Chongqing, China.