Human Action Recognition Based on State Detection in Low-resolution Infrared Video

Tianfu Li, Bo Yang* and Tong Zhang
School of Automation Science and Electrical Engineering
Beihang University
Beijing, China
E-mail: boyang@buaa.edu.cn

2021 IEEE 16th Conference on Industrial Electronics and Applications (ICIEA) | 978-1-6654-2248-2/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICIEA51954.2021.9516410

***Abstract*—This paper proposes a recognition method of human action based on the states in infrared image frames of thermopile array sensors. Two sensors are used for the top-side deployment. Each frame data collected by the thermopile array sensor is called infrared heat map with a resolution 24×32. It is sequentially processed by the quantification, time-domain filtering and background removal operations. Then those processed data are used for the training of state detection models. The state detection models detect the state of the human target in each of frame, and the state sequence is obtained. Finally, A semantic analysis method functions on the state sequence and converts it into the corresponding action. The experimental results have shown that the proposed method can recognize human actions correctly.**

Keywords—Human action recognition; semantic analysis; Infrared heat map; Thermopile array sensor

I. Introduction

In recent years, the word intelligence has frequently appeared in people's vision. Smart phones and smart homes are becoming more and more popular. Smart devices are gradually changing our lives, making our lives more intelligent [1]. Houses are the foundation of daily life, and building smart houses can provide us with great convenience and security. The aging population is becoming an important application of smart houses. It is estimated that by 2053, the proportion of China's elderly population will increase from 16.1% in 2015 to 34.8% [2]. Faced with the large number of elderly populations that is coming, it is difficult to provide perfect care for every elderly person only by relying on labor. Therefore, it is very necessary to build an efficient intelligent care system. The premise of intelligent care is intelligent perception, and the elderly live in indoor environments most of the time. The smart house can provide timely and complete care for the elderly after a comprehensive perception of the behavior.

The premise of building a smart house is to build a fully functional perception system. The usual method is to use the video camera to perceive the situation in the room in real time. However, due to the privacy of the occupant, it may cause discomfort to the occupants, and the video camera system is also affected by the lighting conditions. In addition, the energy consumption of video cameras is relatively large. Through investigation, it is found that the energy consumption of ordinary cameras is at least 100 times that of ordinary low-resolution thermopile infrared array sensors. Considering the popularity of smart houses in the future, the energy consumption of the perception system cannot be ignored. The use of low-resolution infrared sensors can relatively solve the above problems.

At present, there are two kinds of low-resolution infrared sensors in the related research of indoor human target detection: one is pyroelectric infrared sensors (PIS) [3], [4], whose physical characteristics determine that it can only be used to detect moving human targets, and unable to detect stationary targets; the second is the thermopile infrared array sensor (TPAS), which can directly measure the temperature distribution within its field of view. It can detect static and dynamic human targets, providing us with a sense of the environment and the necessary information. In addition, low-resolution thermopile infrared sensors have the advantages of small size, low cost and easy installation. Therefore, there are many researches on using them for indoor human target detection. It mainly includes the number detection, positioning [5] and tracking [6] and behavior recognition of indoor human targets. These can be applied to the perception system of smart houses to realize the most basic perception of human goals.

Taniguchi et.al [7] calculated two features of the thermal data to judge the state of human and designed a sequential transformation rules to determine the behavior. Fan et.al [8] compared the application of several classification deep learning based methods such as Multi-Layer Perceptron (MLP), Long Short Term Memory (LSTM), and Gated Recurrent Unit (GRU) in detecting human falls. Hayashida et.al [9] proposed a method that can detect the falling behavior of human targets in real time in a noisy environment based on a series of logics and thresholds. Mashiyama et.al [10] used Grid-Eye low-resolution thermopile infrared array sensor to recognize five kinds of scenes. The mean-square difference was firstly used to judge whether human behavior is occurring. Then, four features extracted from the data of TPAS will be put into Support Vector Machine (SVM) to classify the specific behavior. Tao et.al [11] used 8×8 infrared array sensors to identify seven daily activities based on a discrete cosine transform. Chen et.al [12] extracted horizontal and vertical features (mean variation, mean absolute deviation, standard deviation of human coordinate) and used K-Nearest Neighbor (KNN) algorithm to detect the falling situation. Since these methods are limited by the resolution of the sensor, the recognized actions are relatively simple.

This paper presents an action perception method based on state detection (see in Fig. 1). It aims to first detect the state of the human target in each frame, and then convert the state sequence obtained by the detection into the corresponding action. The states include standing, sitting, lying, squatting and

978-1-6654-2248-2/21/$31.00 ©2021 IEEE

bending. The corresponding actions mainly include standing, sitting, bending, walking, falling, etc. This paper is organized as follows. Section II introduces the recognition method. The experimental results are introduced in Section III. And finally, Section IV gives the conclusions of this paper.

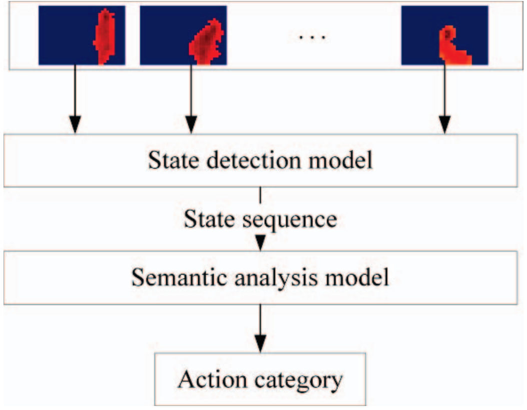


Fig. 1. Human action recognition method

II. Human Action Recognition Method Based on the States

A. Thermopile Array Sensor

Different sensors are applicable to different tasks and scenarios due to their different measurement ranges, measurement accuracy and other parameters. Indoor human target detection and behavior recognition have certain requirements on sensor parameters.

TABLE I
Specifications of MLX90640

Part number	MLX90640ESFBAA
Temperature range	-40~85°C
Resolution	0.1°C
Accuracy	Typ. ±2°C
Number of pixels	24×32
View angle	Typ. 75°×110°
Frame rate	Typ. 8 frames/sec

This paper adopts the thermopile infrared array sensor of MLX90640 (Table I and Fig. 2. (a)), and deploys two sensors on a tripod at the same time, as shown in Fig. 2. (b), to realize the collection of side-view information and top-view information of human targets.

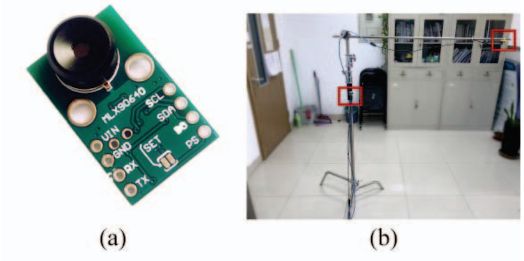


Fig. 2. MLX90640: (a) Sensor (b) Deployment tripod (the red box is the sensor)

The top-view sensor is deployed at a distance of 3m from the ground (sensor 1 in Fig. 3), and the distance between the side-view sensor (sensor 2 in Fig. 3) and the ground is 0.85m. Combining the two sensors, the final effective sensing area is shown in Fig 3, the red cuboid, and the infrared heat map obtained by the collection is shown in Fig. 4.

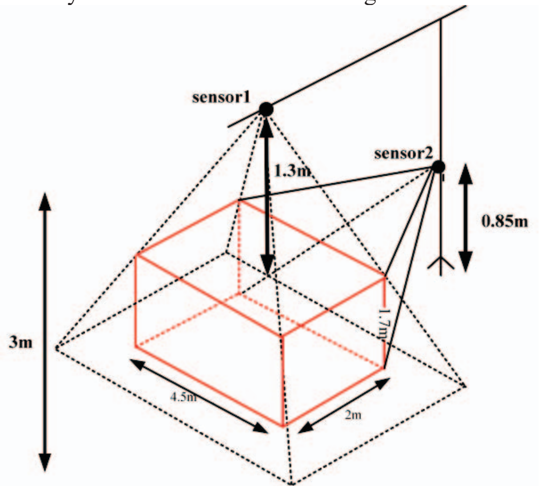


Fig. 3. Effective sensing area (in the red box)

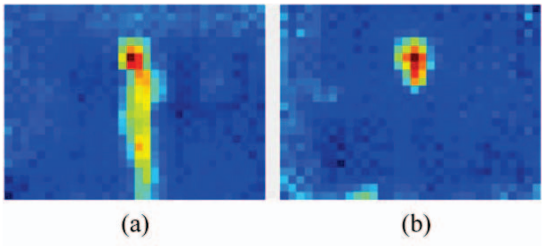


Fig. 4. Infrared heat map of a standing human target: (a) Side; (b) Top

B. Data Processing

The data collected by the thermopile infrared array sensor cannot be directly used as the data for model training, because it contains background information and noise interference. The model will learn this information, thus losing its versatility in different environments and under different interferences. Adolf et.al [13] tried to directly use the unprocessed images obtained by the sensor to train the deep neural network, but the results were not expected, which indicates that image preprocessing is very necessary.

1) Quantification and time-domain filtering

Firstly, the collected temperature value is quantified, and its variation range is extended to 0-255 expressed as:

$$I_t(x,y,t)=\begin{cases} 0 & I(x,y,t)<0 \\ \left\lfloor \dfrac{I(x,y,t)}{0.2} \right\rfloor & 0\le I(x,y,t)\le 51 \\ 255 & I(x,y,t)>51 \end{cases} \quad (1)$$

Where $I_t(x,y,t)$ represents the quantified value of temperature at position (x,y) and time t. $I(x,y,t)$ represents the unquantified value. $\lfloor \bullet \rfloor$ represents round-down operation. The quantified value of temperature is in the variation range of the bit value in the normal picture, which is convenient for using related image processing technology.

Then, the image is filtered by a Gaussian kernel in time domain. The process of filtering is expressed as:

$$I_f(x,y,t)=\sum_{i=t-\tau}^{t} I_t(x,y,i)\cdot k(i) \tag{2}$$

Where $I_f(x,y,t)$ represents the filtered value at position (x,y) and time t . $I_t(x,y,i)$ represents the unfiltered value at position (x,y) and time i . $k(i)$ is the Gaussian kernel and $\tau=3$ is the half of filter window length.

2) Background removal

Due to indoor environment, the interference of the environment is relatively small, and no complex adaptive background removal method is used. Instead, in the absence of a human target, 50 frames of images are continuously collected, and the average of these 50 frames of images is used as background information, and the background information is quantized and time-domain filtered. Finally, the foreground information is obtained by subtracting the background.

As shown in Fig. 5, after the processing of data, the foreground image is the data we finally use to train the model.

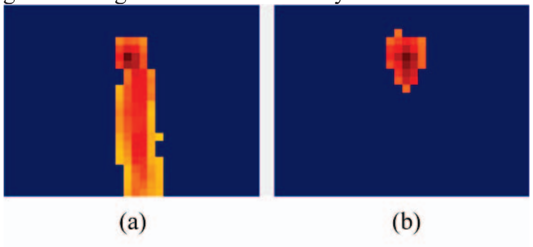


Fig. 5. Infrared heat map after data processing: (a) Side; (b) Top

C. State detection of each frame

Each action of the human target is composed of a continuous sequence of states, so it is necessary to determine the state of the human target in each frame before determining the action. There are mainly five states that need to be recognized: standing, sitting, lying, squatting, and bending. This is a multi-classification problem. We compared two types of detection methods: traditional machine learning methods and deep learning methods.

1) Traditional machine learning methods

Before the training of traditional machine learning model, we should do the feature selections. This paper takes the large differences in the distribution of the five states as the standard, as shown in Fig.6, where in Fig. 6. (a) the five states have large differences, which can be used as the characteristics to distinguish these five states. On the contrary, in Fig. 6. (b), the distribution of the five states is not distinguishable and needs to be discarded. Finally, seven features with large distribution differences were determined. These features need to be further selected based on the training results.

The above seven features are initially determined by analyzing the difference in feature distribution on the data set. The following training SVM and MLP classifiers use the average cross-validation accuracy rate on the test set as an indicator to finally determine the features used for training. We take the center of gravityg_y (side view),variance h(side view), variance w(top view), h as the basic features. Because the difference in the distribution of these four features is relatively largest, and if only a few features are used, the generalization ability of the model will be reduced.

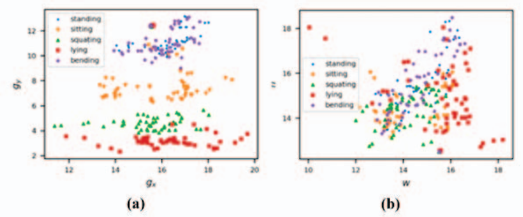


Fig. 6. A comparison of feature distributions (center of gravity): (a) Side; (b) Top

The data set contains a total of 251 samples on 5 states of 6 human targets with different orientations (relative to the position of the side-view sensor). Label six human targets as 1, 2, 3, 4, 5, and 6. The cross-validation method is used for the training. The accuracy is the average accuracy on the test set. The construction of the model can be directly called using Python's sklearn library.

TABLE II
SVM TRAINING RESULTS USING DIFFERENT FEATURES

Selected features	Average accuracy
Basic features	86.04%
Basic features + Orientation	84.86%
Basic features + Eccentricity	85.64%
Basic features + Equivalent diameter	81.28%
Basic features + Orientation + Equivalent diameter + Equivalent diameter	81.67%

Table II shows the results of training using these four basic features. Its average accuracy on the test set is 86.04%. Table 2 shows that the newly added features do not improve the accuracy of the SVM classifier. To eliminate the tendency of the classifier, we chose a multi-layer perceptron (MLP) to train again. The training results are shown in Table III. The newly added feature long axis orientation and equivalent diameter have a certain improvement in its classification accuracy, but the improvement is relatively small, and these newly added features are not key features.

TABLE III
MLP TRAINING RESULTS USING DIFFERENT FEATURES

Selected features	Average accuracy
Basic features	81.69%
Basic features + Orientation	82.10%
Basic features + Eccentricity	81.70%
Basic features + Equivalent diameter	82.49%
Basic features + Orientation + Equivalent diameter + Equivalent diameter	80.90%

Using all the features to train the model does not improve the accuracy of the model, on the contrary its accuracy is the worst. The reason may be that the introduction of certain features makes the feature distribution of the sample greatly changed, making it difficult for the classifier to make reasonable divisions. As shown in Fig. 7, when new features are added, the feature distribution becomes different, and the corresponding classifier type should also change, that is, from a green straight line to a red polyline.

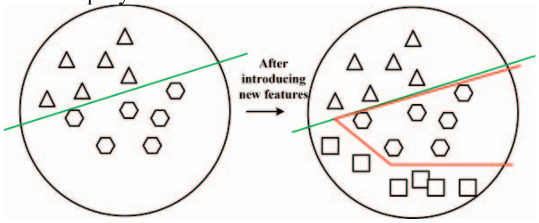


Fig. 7. The changes of classifiers with the addition of new features

2) Deep learning methods

The classification accuracy of traditional classifiers is greatly affected by the selected features. Through these limited characteristics, it is difficult to distinguish relatively harsh and complex transition states. Convolutional neural network (CNN) is commonly used to extract features of two-dimensional image data. It continuously performs feature extraction and compression by continuously sliding its convolution kernel and performing convolution operations with the area covered by the convolution kernel. The final feature vector is the feature of the entire image information, not just the extraction of some local features like the traditional classifier. Therefore, the characteristic of the convolutional neural network is that it can automatically extract the features, and it is more comprehensive and richer than the artificial selection.

The paper uses keras deep learning framework to build the network structure. The network for one view includes three convolutional layers and three pooling layers, a fully connected layer for feature fusion, and finally a softmax layer to give the classification result .

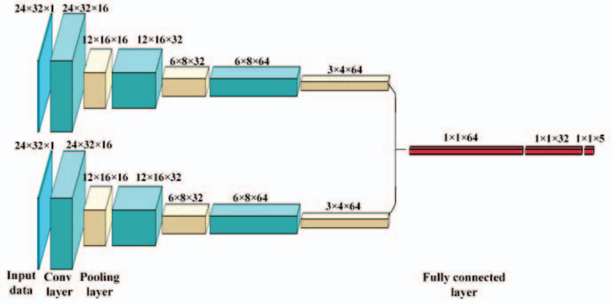


Fig. 8. The convolutional neural network structure using both top and side view data as input data.

Since we have two sets of data, including side-view sensor and top-view sensor data, we use two convolutional neural networks to extract the features of the two sets of data respectively, and then combine the two sets of features, and finally pass to the fully connected layer and Softmax for completing the classification (see in Fig. 8). The training results are shown in Table IV.

TABLE IV
TRAINING RESULTS ON DIFFERENT DATA SETS

Training data	Average accuracy
Side view data	95.85%
Top and side view data	96.73%

D. Semantic analysis

The main purpose of semantic analysis is to transform the sequence of states into corresponding actions, and at the same time use the continuity of actions to correct some false detections. There are two main reasons why the detected state of each frame cannot be directly used as the result: one is that the model is not very accurate in determining the transition state between successive actions, and the other is that it needs to distinguish between walking and standing close actions.

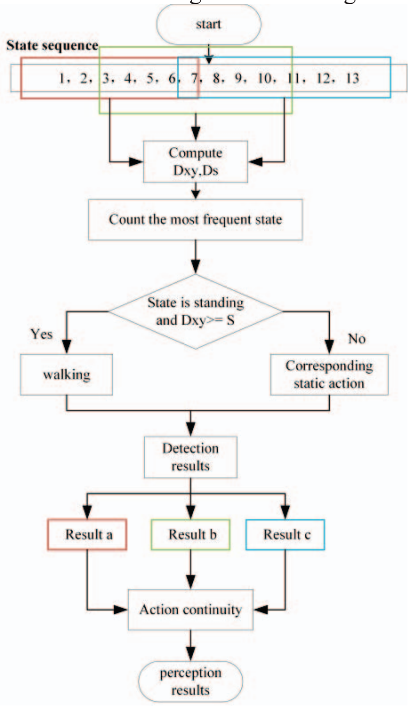


Fig. 9. Action semantic analysis algorithm

We define a buffer of length L, which stores L frames of continuous infrared heat maps. As the detection progresses, the first frame is continuously deleted and the current frame is placed in the buffer. The infrared heat map is transformed into a state sequence labeled from 1 to L through the state detection model. Then we divide the L continuous states into three parts, and the two adjacent parts can cross each other. For each part of the infrared heat map, we calculate the distance moved by the center of gravity of the top-view infrared heat map and count the number of occurrences of each state. If the main state of the human target in this part is standing, and the distance D_{xy} moved by the top-view center of gravity exceeds the threshold S, then we determine that the human target is walking at this time. If the threshold is not exceeded or the state is not standing, then the determination result is the main status in this section. The first part of the behavioral perception result is named Action a, and the second and third part of the results are Action b and Action c, respectively. Finally, it is judged by the continuity of the action, the purpose is to determine the category of Action b, which is determined as the judgment result of the current human target behavior. If Action a and A ction c are the same, then Action b must be the same as Actions a and c, because there is no time interval between Actions a and c, and b cannot be another action. If a and c are different, then b must be one of the two actions, because it contains both a part of a and c, so if b and a are of the same type, if the frame of action a count more, so return to action a, and vice versa for action c. Table V shows the pseudo code for determining the action continuity. Finally, the complete semantic analysis process is shown in Fig. 9.

In the above method, two layers of semantic analysis are used. The first layer determines the corresponding actions by analyzing the continuous state. However, in the continuous action recognition test, it is found that the stability of the semantic analysis of the first layer is not good enough. The second level of semantic analysis is to analyze the detection results of the three parts, and use the continuity criterion of the action to determine the action category of the middle part. The length of the buffer area L will affect the detection speed and detection delay. We use the detection result in the middle part as the current detection result, and the delay is about L/4~L/2 frames. The acquisition rate of the sensor is 8 frames/s, so the delay of the semantic analysis process is between L/32~L/16s. If L is too large, the delay will be too long. If L is too small, the value of semantic analysis will be lost and the stability of the detection result will not be improved. Finally, through preliminary experimental tests, L is determined to be 13, and the delay is between 0.41 and 0.81s.

TABLE V
PSEUDO CODE BASED ON ACTION CONTINUITY

```
Input: Three consecutive detected actions: a, b, c;
a has been corrected;
Output: Type of action b after correction
if a == c:
        return a
if a == b:
        return a
else:
    return c
```

III. EXPERIMENTAL RESULTS

A. Setup of Experiment

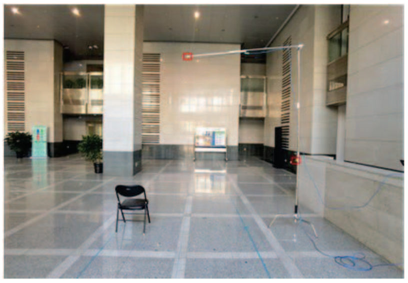
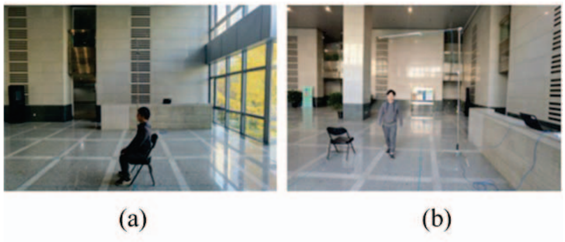


Fig. 10. Experimental test scenario



(a) (b)

Fig. 11. Some experimental test actions: (a) Sitting; (b) Walking

The paper designs a comprehensive test experiment of the algorithm in actual indoor scenarios. The experimental site is shown in Fig. 10. The red box is the location of the two sensors. Fig. 11 shows some experimental action test diagrams.

B. Results of Experiment

1) Continuous action test

The detection of continuous actions mainly tests the accuracy and delay of the algorithm, and whether it can effectively distinguish the transition phase between two actions. The paper designs a set of continuous actions as shown in Fig. 12 for testing. It mainly includes four movements: walking, bending, sitting, and squatting.

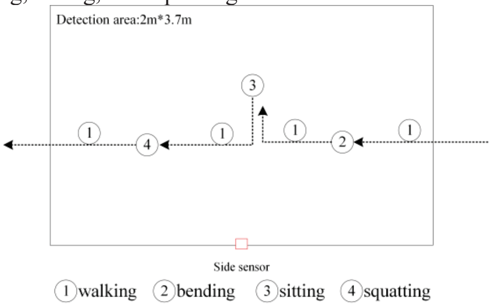


Fig. 12. Design of continuous action test

During the experiment, the human target completes the actions in Fig. 12 at a normal speed. The comparison between the detection result and the actual result is shown in Fig. 13. The actual result is represented by a red line, the detection result is represented by a blue dashed line, and the abscissa is the detected frame. The ordinate is the behavior category of the human target. We can first see that there is a delay of about 4 frames between the detection and the actual result. Except for a slight fluctuation at the end of the bending motion, the rest of the detection results are relatively accurate.

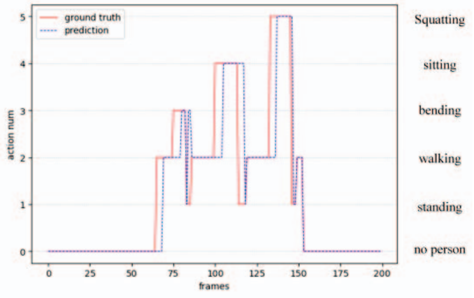


Fig. 13. Comparison of test results with ground truth

Since the acquisition frequency of the sensor is 8 frames per second, the delay of 4 frames is about 0.5 seconds. In the Section II, we analyzed that the delay of the entire algorithm is probably between L/32s~L/16s. In the actual experiment, we take 13 frames for L, that is, the delay is between 0.4s~0.81s. It is quite consistent with our actual test results. In addition, we will focus on the division of transition actions between the two actions. In the division of the transition state of the actual result, as long as the second action has a trace of the beginning, it is divided into the next action. Such a division is difficult for the detection result to match the actual result, because when the action changes in the beginning, the information retained in the buffer is still in the main part of the previous action, and the new action has less information. Only when the new action has relatively more information can the detection result change.

2) An attempt to detect the edge action

The detection of sensor edge motion is a more challenging task. There are two main factors that affect the final detection result:

1) Sensor detection accuracy: The sensor used in the paper experiment has a resolution of only 24×32. When the human target is at the edge, some information will be lost, which will increase the difficulty of detection.
2) Distortion of edge position: All cameras have distortion, and the degree of distortion increases as the proportion of the field of view corresponding to each pixel increases. Since the thermopile infrared array sensor, we use is equivalent to a low-resolution infrared camera, due to its own viewing angle and other factors, there will be a large distortion at the edge, which may affect the detection result.

Due to the above two problems, the shape of the infrared heat map of the human target at the edge position is different from other positions. Therefore, the detection result of the method of using traditional manual feature extraction for recognition will be poor. This is also why the deep learning method is used in the paper. The paper experiment mainly tested the positions of the four corners, as shown in Fig. 14. The experiment tested three movements of bending, sitting and squatting in four positions. The experimental results are shown in Table VI. A tick indicates that it can be recognized correctly, and a cross indicates that the corresponding action cannot be correctly recognized.

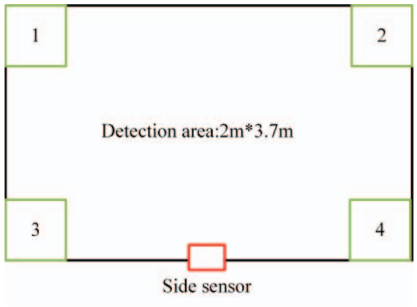


Fig. 14. Four positions for edge area testing (from top view)

From the detection results in Table VI, the algorithm can correctly identify the two actions of squatting and sitting even at the edge position. Because these two actions are relatively different from other actions, they can be detected even if some information is lost. On the contrary, for standing and bending, the key feature of the two actions is considered to be the height of the human target. When the human target is at the edge, the infrared heat map of the human target becomes larger when the side-view sensor is closer, the human target becomes relatively smaller when it is far away from the side-view sensor. The enlargement and reduction of human targets will affect the results of the detection.

TABLE VI
EDGE AREA ACTION TEST RESULTS

Position	Standing	Bending	Squatting	Sitting
1	×	×	√	√
2	×	√	√	√
3	√	√	√	√
4	×	√	√	√

IV. CONCLUSION

This paper presents an action perception method based on state detection. It aims to first detect the state of the human target in each frame, and then convert the state sequence obtained by the detection into the corresponding action. The state detection method use the deep learning method which extracts automatically features and performs better than the traditional machine learning methods.

V. ACKNOWLEDGMENTS

The authors would like to express their gratitude for the financial support by National Natural Science Foundation of China (No. 61773034).

REFERENCES

[1] Wang Qing. Research on the construction and application of a new smart house system[J]. China New Technology and New Products, 2018(8):22-23
[2] Jiang Chunli. Analysis of the status quo of my country's population aging and countermeasures and strategies during the "13th Five-Year Plan" period[J]. Globalization, 2016(8):90-105
[3] S. Lee, K. N. Ha, and K. C. Lee. “A pyroelectric infrared sensor-based indoor location-aware system for the smart home.” *Consumer Electronics IEEE Transactions on* 52(2006):1311-1317.
[4] S. O. Al-Jazzar, S. A. Aldalahmeh, D. McLernon and S. A. R. Zaidi, “Intruder Localization and Tracking Using Two Pyroelectric Infrared Sensors.” *IEEE Sensors Journal*, vol. 20, no. 11, pp. 6075-6082, 1 June1, 2020, doi: 10.1109/JSEN.2020.2974633.
[5] M. Kuki, H. Nakajima, N. Tsuchiya and Y. Hata, “Multi-Human Locating in Real Environment by Thermal Sensor,” in *Proc. SMC* 2013, Manchester, United Kingdom, 2013, pp. 4623-4628.
[6] A. D. Shetty, Disha, Shubha, B. and Suryanarayana, K, “Detection and tracking of a human using the infrared thermopile array sensor- ‘Grid-EYE’,” in *Proc. ICICICT*, Kerala State, Kannur, India, 2017, pp. 1490-1495.
[7] Y. Taniguchi, H. Nakajima, N. Tsuchiya, J. Tanaka, F. Aita and Y. Hata, “A falling detection system with plural thermal array sensors,” in *Proc. SCIS&ISIS*, Kitakyushu, Japan, 2014, pp. 673-678.
[8] X. Fan, H. Zhang, C. Leung, Z. Shen, “Robust unobtrusive fall detection using infrared array sensors,” in *Proc. MFI* 2017, Daegu, Korea, 2017, pp. 194-199.
[9] A. Hayashida, V. Moshnyaga and K. Hashimoto, “The use of thermal ir array sensor for indoor fall detection,” in *Proc. SMC*, 2017, pp. 594-599, doi: 10.1109/SMC.2017.8122671.
[10] S. Mashiyama, J. Hong and T. Ohtsuki, “Activity recognition using low resolution infrared array sensor,” in *Proc. ICC* 2015, London, United Kingdom, 2015, pp. 495-500.
[11] L. Tao, T. Volonakis, B. Tan, et al, “Home Activity Monitoring using Low Resolution Infrared Sensor,”. arXiv preprint arXiv:1811.05416, 2018
[12] W. H. Chen and H. P. Ma, “A fall detection system based on infrared array sensors with tracking capability for the elderly at home,” in *Proc. HealthCom*, Boston, USA, 2015, pp. 428-434.
[13] J. Adolf, M. Macas, L. Lhotska and J. Dolezal, “Deep neural network based body posture recognitions and fall detection from low resolution infrared array sensor,” in *Proc. BIBM*, 2018, pp. 2394-2399, doi: 10.1109/BIBM.2018.8621582.