Robust Activity Recognition for Aging Society

著者	CHEN Yi, YU Li, OTA Kaoru, DONG Mianxiong
journal or publication title	IEEE journal of biomedical and health informatics
volume	22
number	6
page range	1754-1764
year	2018-03-26
URL	http://hdl.handle.net/10258/00009931

doi: info:doi/10.1109/JBHI.2018.2819182

Robust Activity Recognition for Aging Society

Yi Chen, Li Yu, *Senior Member, IEEE* Kaoru Ota, *Member, IEEE,* and Mianxiong Dong *Member, IEEE,*

***Abstract*—Human activity recognition(HAR) is widely applied to many industrial applications. In the context of Industry 4.0, driven by the same demand of machines' self-organizing ability, HAR can be also adopted in elderly healthcare. However, HAR should be adaptive to the application scenarios in elderly healthcare. In this paper, we propose a non-intrusive activity recognition method which can be applied to long-term and unobtrusive monitoring for elderlies. The method is robust to obstruction and non-target object interference. Skeleton sequence is estimated from RGB images. Based on two activity continuity metrics, an Inter-frame Matching Algorithm is proposed to filter non-target objects. In order to make full use of spatial-temporal information, we propose a novel activity encoding method based on the interframe joints distances. A convolutional neural network is used to learn the distinguishing features automatically. A specific data augmentation method is designed to avoid the over-fitting problem on small-scale datasets. The experiments are performed on two public activity datasets and a newly released Noisy Activity Dataset(NAD). The NAD contains obstruction, non-target object interference. The experimental results show that the proposed method achieves the state-of-art performance while only using one ordinary camera. The proposed method is robust to a realistic environment.**

***Index Terms*—Industry 4.0, Human activity recognition, Elderly healthcare, Unobtrusive monitoring, CNN.**

I. INTRODUCTION

With the rapid development of Industry 4.0, the demands for the human and machine are also changing dramatically [1]. In order to make services and production more intelligent, machines including robots are required to possess autonomy and self-organizing ability [2]. Drove by this request, human activity recognition is widely applied to various industrial applications, including work analysis of factory workers [3] and production monitoring [4].

In the context of industry 4.0, various automation technologies including human activity recognition are also applied to many healthcare applications such as home-oriented health monitoring service [5], long-term monitoring for elderly [6], remote medical [7], and primary care [8]. Although human activity recognition is widely used in healthcare, there are still some different demands when compared with industrial applications. Different from the necessary cooperation between factory workers and machines or robots, it is not reasonable to ask the target object for cooperation especially in some elderly healthcare applications [9]. For example, sensor based activity recognition methods [10] may cause living inconvenience for elderly. Thus, it is necessary to develop unobtrusive activity recognition method for many elderly healthcare applications such as long-term monitoring for elderly, and remote medical.

Most current research work focuses on achieving high recognition accuracy, which may be the most important condition. The previous research tried to accomplish activity recognition based on video sequences captured by common RGB video cameras [11]–[15]. However, these methods needed to obtain adequate silhouette features. In order to extract silhouette features, the complex and time-wasting processing chain (e.g., background removal, vector quantization, image normalization) was also necessary, which limited the real-time application.

The high time complexity and unsatisfactory recognition accuracy are two general weakness of early research based on RGB data, which forced most researchers to use other types of activity data. Wearable sensors and depth camera are two preferred types of data capture equipment. However, there are some factors which limit the large-scale promotion of such equipments in elderly healthcare.

Many wearable sensors were based on accelerometry such as wrist-worn or waist worn devices. Wrist-worn devices may perform poorly on lower body activities classification [16] [17]. Waist worn accelerometers may make upper body activities undetected [18]. To overcome these shortcomings, some researchers proposed to utilize multiple sensors [19]–[23] or the combination of wearable sensors and other devices(cameras [24], [25], smartphone [26], wireless [27]). Multiple wearable sensors based recognition systems are sensitive to each sensor. If some sensors are unable to capture and transmit data normally, these methods' accuracy may be badly influenced. Additionally, intrusive wearable sensors could cause inconvenience for the daily life of target objects.

As for depth camera, except for RGB information, the depth information of the corresponding activity scene can be also provided. Depth maps captured by depth cameras were applied to estimate 3-D posture for activity recognition [28]–[31]. The quality of depth map plays an important role. However, limited by the hardware of depth cameras, depth map contains strong noise as well as holes [32], [33]. Additionally, its resolution is lower than the corresponding RGB image. These factors will influence the estimation of 3-D posture. Thus, RGB-D data based recognition systems are not ideal choice for large-scale promotion neither.

With the improvement of depth sensors, nowadays, some sensors such as Prime Sensor [34] can provide the real time 3D joints information of the target object when the target object accomplishes a calibration pose facing the camera. These 3D joints information is also called skeleton data. Compared with

Y. Chen, L. Yu are with the Department of Electronics and Information Engineering, Huazhong University of Science and Technology, Wuhan, China. Y. Chen is also an Visiting Student with the Department of Information and Electronic Engineering, Muroran Institute of Technology, Muroran, Hokkaido, Japan.

E-mail: m201671914@hust.edu.cn; hustlyu@hust.edu.cn

K. Ota, M. Dong are with the Department of Information and Electronic Engineering, Muroran Institute of Technology, Muroran, Hokkaido,Japan. M. Dong is the corresponding author.

E-mail: ota@csse.muroran-it.ac.jp; mx.dong@csse.muroran-it.ac.jp

other types of activity data, skeleton data is more related to human activity. Additionally, skeleton data can reduce feature extraction time and provide adequate accurate features. Thus, skeleton data based recognition systems [35]–[37] showed more outstanding performance. However, there are three basic assumptions for these skeleton tracking sensors:(1)User's upper body is mostly inside the field of view. (2)Ideal distance is around 2.5 m. (3)For better results, the user should not wear very loose clothing. If the three assumptions are not satisfied, Some key body points may be undetected. As far as we know, there are no activity recognition methods using incomplete skeleton data. In a word, calibration pose and three basic assumptions are the factors which limit the large-scale promotion of these recognition systems. But these methods still showed the superiority of skeletal data.

In this paper, our goal is to design a robust, unobtrusive activity recognition method which can be applied to elderly healthcaer applications such as long-term monitoring for elderly, and remote medical. Based on the analysis above, our idea is to capture RGB images using an ordinary camera. Then, existed posture estimation algorithm is applied to estimate skeleton sequence from RGB images. In order to avoid overlooking any important information, we choose to preserve all the spatial-temporal information in our activity encoding method. Then a convolution neural network is used to learn those distinguishing features automatically. Since convolution neural networks are prone to overfit on small-scale datasets, a specific data augmentation algorithm is designed. Additionally, obstruction and non-target objects interference are also considered in our method. Although there are no public datasets which contain obstruction and non-target objects interference, it can be solved by collecting a new dataset. Thus, the contributions of this paper can be stated as follows:

(1) A robust, unobtrusive activity recognition method based on RGB images is proposed. The proposed method can achieve the state-of-art accuracy while using less information.

(2) An Interframe Matching Algorithm is proposed to filter non-target objects of skeleton sequence. Since the actions of the same person in the adjacent two frames will not change drastically, we defined two types of continuity metrics. The continuity metrics help us extract skeleton sequence of the target object.

(3) A novel activity encoding method is proposed to preserve both spatial and temporal information. Based on joints distances of two skeletons in the adjacent two frames, each skeleton is converted to a feature vector. Thus, all the information of the skeleton sequence is preserved. The encoding method is valid to incomplete skeleton data.

(4) A specific data augmentation algorithm is designed to generate adequate training samples. Since activity is continuous, increasing the sampling frequency won't change the activity itself. Thus, we inserted multiple skeletons in the adjacent two frames firstly. Then, multiple samples are generated by extracting a fixed amount of skeletons evenly.

(5) we release a noisy activity dataset which contains obstruction and non-target objects interference. Since other dataset doesn't consider these factors, our dataset is more close to the realistic environment. It can help assess the robustness of the proposed method.

This paper is organized as follows. Pose estimation algorithm and activity datasets are outlined firstly. Next, the proposed activity recognition method is described. Then the experimental results are presented. Finally, we will discuss about our method and make a conclusion.

II. Related Work and Datasets

Firstly, we review 2D pose estimation algorithm. Then, a brief description of the two public activity datasets will be provided. Finally, the details of the noisy activity dataset are presented.

A. 2-D Pose Estimation

Pose estimation can be regarded as a subtask in human activity recognition. Most pose estimation methods [38] [39] can be divided into two steps: person recognition and location, key body points detection. These methods are limited by the complex process (e.g. background removal, person detection). These factors are also the reasons why some researchers gave up activity recognition based on RGB images stream. In [40], Cao proposed an algorithm which can achieve real-time multi-person 2D pose estimation using part affinity fields. The algorithm detected key body points and calculated the optimal connection of all the key body points at the same time. In this paper, the algorithm [40] is used to estimate skeleton data. In fact, any posture estimation algorithms which can satisfy the real-time requirement is acceptable for our methods.

The skeleton data extracted by [40] is a little different from the skeleton data captured by skeleton tracing sensors. Firstly, Skeleton data captured by skeleton tracing sensors contains less noise when the three basic assumptions are satisfied. Each key body point's information is complete. Skeleton data generated using estimation algorithm may be incomplete due to obstructions. Secondly, skeleton tracing sensors can only provide the skeleton data of the target object. Posture estimation algorithm can estimate skeleton data of all the potential persons. Skeleton tracing sensors were suitable to be applied in some specific application such as virtual reality [41] and human-computer interaction game [42], since these application scenes contains little or no noise (eg. obstructions, non-target objects). Pose estimation algorithms are more suitable to capture skeleton data in the real environment, since the target object won't always accomplish the calibration action in their real life.

B. Public Activity Datasets

In this paper, we will perform experiments on two public activity datasets: Kinect Activity Recognition Dataset(**KARD**) [28], **Florence 3D Action** [43]. The details of the two datasets are described respectively.

a) **KARD:** This dataset is collected in 2015. It contains ten gestures(hand clap, bend, sidekick, forward kick, draw tick, draw x, high throw, two hands wave, high arm wave, and horizontal arm wave) and eight actions(stand up, sit down, phone call drink, walk, take umbrella, toss paper, and catch

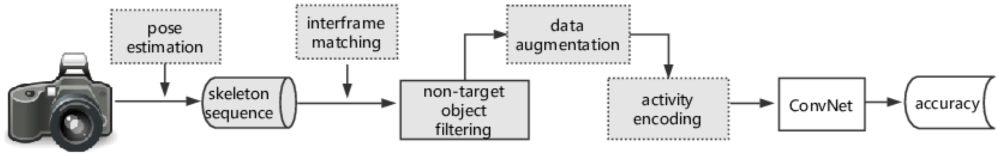


Fig. 1. The scheme of the proposed method. RGB images captured by camera are processed to generate skeleton sequence using posture estimation algorithm. Then the interframe matching algorithm is applied to filter non-target objects. In training, adequate samples are generated using a new data augmentation algorithm. At last, skeleton sequence is encoded and fed into a ConvNet for recognition

up). Each activity is performed three times by ten different persons. Thus, there are 540 sequences of videos.

b) **Florence 3D Action:** This dataset is collected at the university of florence using a Kinect camera. It contains 9 activities(drink from a bottle, arm wave, sit down, clap, answer phone, tight lace, bow, stand up, read watch). Each activity is repeated by 10 persons several times for a total of 215 sequences.

There are some common features in these two public datasets. Firstly, the two datasets provide RGB-D data and skeleton sequences. Secondly, there are no obstructions during the process of data collection. The body of the target object is always visible. Thirdly, skeleton sequence only contains the target object. Although there are non-target objects in some videos, the non-target objects don't interfere with the activity of the target object. Additionally, the target object is always facing the camera. In this paper, we only utilize the RGB images of these two datasets.

C. Noisy Activity Dataset

Most current public datasets are similar to the two datasets above. There are no obstructions resulted from obstacles or non-target objects. It's not consistent with the real environment. Thus, we collect a noisy human activity dataset which considers both obstructions and non-target objects. In this noisy dataset, the part body of the target object may be invisible. Additionally, non-target objects may interfere with the activity of the target object. The dataset is denoted as **Noisy Activity Dataset**(**NAD**[1]).

Our dataset contains 7 daily activities: Bow, Walking, Playing tennis, bending and picking up items, taking shoulder bag, walking while pushing chair, bending and tying a shoelace. Each activity is performed 3 times by 4 persons for a total of 84 sequences. Each sequence is recorded using an ordinary camera. The target object is only requested to perform activities without any special requirements, such as always facing the camera. For each activity, there are no noises(obstructions, non-target objects) when each person performs each activity the first time. When the target object performs the same activity the second time, there are some obstructions such as desk, being placed between the target object and the camera. These obstructions make a part body of the target object invisible. When the target object performs the same activity the third time, the non-target objects and obstructions exists at the same time. The non-target objects enters the field of view at some time point in the middle. Furthermore, the non-target objects do anything (walking through between the camera and the target object, entering and doing the same activity) randomly.

Except for the obstruction and non-target objects, our dataset also keeps the similarity of different activities. In this paper, the experiments on our **Noisy Human Activity Dataset** are also performed based on the RGB images.

[1]The Noisy Activity Dataset can be downloaded via https://github.com/Luyaojun/Noisy-ActionDataset

III. ACTIVITY RECOGNITION USING RGB IMAGES

The proposed method shown in Fig. 1 aims at achieving high-precision activity recognition using RGB images. There are four key components in our method. Firstly, the non-target objects of the skeleton sequence are filtered using **Interframe Matching Algorithm**. Secondly, a large number of activity samples are expanded with specific data augmentation algorithm. Thirdly, the spatial-temporal information of skeleton sequence is encoded with a novel encoding method based on joints distances in the adjacent two frames. Finally, a convolutional neural network is trained to select the distinguishing features and recognize activities.

A. Skeleton filtering by Interframe Matching Algorithm

Due to obstructions, skeleton data may be incomplete. As shown in Fig.2, the left one is ideal skeleton data, which contains **r** key body points. Each key point is determined by the corresponding coordinate information, which is denoted as (x, y). The right one is incomplete skeleton data. The coordinate of the undetected key points is denoted as $(0, 0)$. Additionally, there may be multiple skeleton data in some frame images.

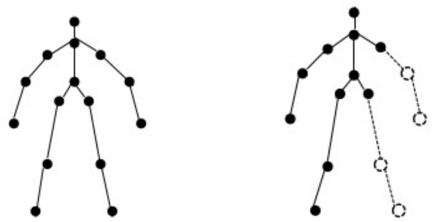
Let's assume that there total **t** frame images in a video. Then the skeleton sequence can be formulated as

$$video = \{F_1, F_2 ... F_t\} \tag{1}$$

$$F_i = \{\overline{P^i_1}, \overline{P^i_2} ... \overline{P^i_{m_i}}\}, i \in [1, t], m_1 = 1 \tag{2}$$

$$\overline{P^i_j} = [x^{i,j}_0, y^{i,j}_0 ... x^{i,j}_{r-1}, y^{i,j}_{r-1}], j \in [1, m_i] \tag{3}$$

where F_i is the i_{th} frame image in the video. $\overline{P^i_j}$ is the j_{th} skeleton data in the i_{th} frame image. x, y are the coordinates of the corresponding key points.



(a) ideal skeleton (b) noisy skeleton

Fig. 2. (a) is the ideal skeleton. (b) is the noisy skeleton. The hallow circle stands for the undetected key points

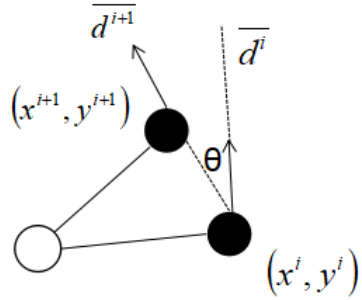


Fig. 3. The two black points are the same key point in the i_{th} and $i+1_{th}$ frame image respectively

In the $i+1_{th}$ frame, both the spatial position and movement direction of each key point won't change violently when compared with the same key point in the i_{th} frame image. Based on these two clues, we can know that there are two types of continuities for each key point between two continuous frame images. The first one is denoted as spatial continuity while the other one is denoted as direction continuity. According to the continuities of all the key points, the similarity between two skeletons can be calculated, which can tell the target object from all the objects. The method is also called **Interframe Matching Algorithm**.

Before we give the detail of the algorithm, the two types of continuity need to be made clear. Fig. 3 shows the same key point in two continuous frame images.

$\overline{d^i}$ and $\overline{d^{i+1}}$ are the corresponding directions. $\left(x^i, y^i\right)$ and $\left(x^{i+1}, y^{i+1}\right)$ are the corresponding coordinates of the black key point. In order to measure the two types of continuity, the similarity of directions in two frame images is regarded as the metrics of direction continuity. The distance between the two coordinates is regarded as the metrics of the spatial continuity. The two metrics are formulated as

$$c_k = \theta = \arccos\left(\frac{\overline{d^i} \bullet \overline{d^{i+1}}}{\left|\overline{d^i}\right| \times \left|\overline{d^{i+1}}\right|}\right), k \in [1, r] \tag{4}$$

$$s_k = \sqrt{\left(x^i - x^{i+1}\right)^2 + \left(y^i - y^{i+1}\right)^2}, k \in [1, r] \tag{5}$$

where S is the metrics of the spatial continuity. C is the metrics of the direction continuity. For the key point $\left(x^{i+1}, y^{i+1}\right)$, its direction is expressed as

$$\overline{d^{i+1}} = \left(x^{i+1} - x^i, y^{i+1} - y^i\right), i \in [1, t-1] \tag{6}$$

For the 1_{th} frame, the directions of all the key points are all initialed as $(0, 0)$.

Once we got the continuities information of each key point, the continuities of the whole skeleton can be also defined. The skeleton is composed of **r** key points. However, the skeleton may not be complete. Thus, the continuities of the whole skeleton are determined by the valid key points which are both detected in two skeletons. Assuming that $\overline{P_j^i}$ is the j_{th} skeleton in i_{th} frame image and $\overline{P_q^{i+1}}$ is the q_{th} skeleton in $i+1_{th}$ frame image, the continuities of the two skeleton is formulated as

$$C_{j,q}^{i,i+1} = \frac{\sum_{k\in\Omega} c_k}{N_\Omega \times \pi}, \qquad S_{j,q}^{i,i+1} = \frac{\sum_{k\in\Omega} s_k}{N_\Omega} \tag{7}$$

$$\Omega \in \{k \left| \left(x_k^{i,j}, y_k^{i,j}\right) \neq (0,0) \wedge \left(x_k^{i+1,q}, y_k^{i+1,q}\right) \neq (0,0)\right\} \tag{8}$$

where Ω is the set of the valid key points in the two skeletons, and N_Ω is the amount of the valid key points.

The direction continuity and spatial continuity of two skeletons measured the similarity of action direction and position respectively. In other words, the probability that two skeletons belong to the same person is depended on the direction continuity and spatial continuity at the same time. Here we proposed a new concept named similarity score F. If F is a large number, the probability is also large. The similarity score is formulated as

$$\begin{aligned} &F_{j,q}^{i,i+1} = \alpha \times e^{-C_{j,q}^{i,i+1}} + \beta \times e^{-S_{j,q}^{i,i+1}} \\ &i \in [1, t-1],\ j \in [1, m_i], q \in [1, m_{i+1}] \end{aligned} \tag{9}$$

where α and β are the weight parameters.

Now it's possible to filter non-target objects in the skeleton sequence. Let's assume that the target object in i_{th} frame image is the j_{th} person P_j^i. If there are m_{i+1} skeletons of the $i+1_{th}$ frame image. Then target object in the $i+1_{th}$ frame image can be found by comparing the similarity scores.

$$\mathbf{tar} = \arg\max_q \left(F_{j,q}^{i,i+1}\right), i \in [1, t-1], q \in [1, m_{i+1}] \tag{10}$$

where **tar** is the index of that target object in the $i+1_{th}$ frame image. In this paper, **Interframe Matching Algorithm** is used to search for the target object in the next frame image. By repeating the process, the complete skeleton sequence of the target object can be extracted from the noisy skeleton sequence.

B. Data Augmentation

An activity is composed of a series of skeleton. The skeleton sequence is different from other types of data such as images. The conventional augmentation technologies including rotation, cropping randomly and so on, are not applicable any

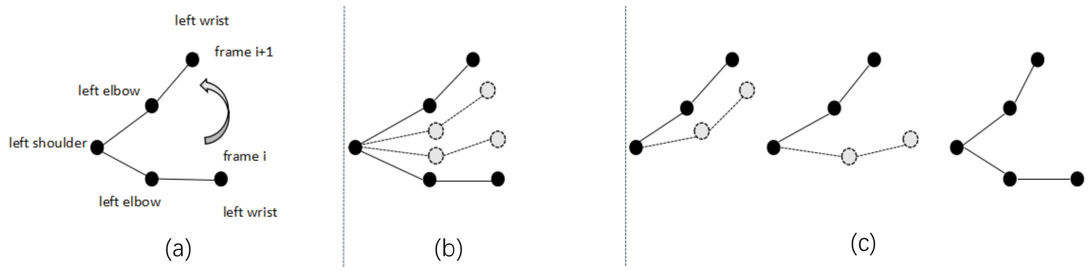


Fig. 4. (a) shows the raw 'raise hand', (b) shows the action after interpolating, (c) shows three 'raise hand' samples after down-sampling

more. To the best of our knowledge, there are no dedicated data augmentation algorithms for skeleton sequence.

Some researchers tried to use conventional augmentation algorithm in another way. In [44], the authors converted a skeleton sequence to a color image. Then they used conventional augmentation technology(cropping randomly, rotation) to expand training samples. Their method can create adequate training samples. However, whether the expanded samples are still the same type of activity is not sure. Since the order of the skeleton in all the frame images is relevant to the specific type of activity, rotation may change the type of activity. As for the process of cropping randomly, it's also not suitable because it may make the cropped activity sample incomplete. The order and completeness of skeletons are also the basis for distinguishing different types of activities.

In this paper, we propose a simple but effective augmentation algorithm for skeleton sequence. Our augmentation algorithm can be divided into two steps: insert skeleton evenly, and extract skeletons using a sliding window. Fig.4 is an example of our augmentation algorithm.

Fig.4(a) shows the raw action of 'raise hand' which contains two skeletons. Fig.4(b) shows the same action after dense interpolating. Although the skeleton sequence is discrete, the movement of the body is continuous. Thus, we can insert several skeletons evenly in two continuous frame images, which is equivalent to increase the frequency of capturing the posture of the target object. Then, we can generate multiple action samples by down-sampling as Fig.4(c)shows. The three samples are not the same. But the three samples represent the same action of 'raise hand'.

Assuming that $\overline{P_j^i}$ and $\overline{P_q^{i+1}}$ are the skeletons of the target object in the i_{th} and $i+1_{th}$ frame image respectively. We need to insert **scale** skeletons between the two frame images. Then the process can be performed as

$$\overline{P_j^i} = [x_0^{i,j}, y_0^{i,j} ... x_{r-1}^{i,j}, y_{r-1}^{i,j}], i \in [1, t-1] \tag{11}$$

$$\overline{P_q^{i+1}} = [x_0^{i+1,q}, y_0^{i+1,q} ... x_{r-1}^{i+1,q}, y_{r-1}^{i+1,q}], q \in [1, m_{i+1}] \tag{12}$$

$$\overline{P_s^{i,i+1}} = [x_{0,s}^{i,i+1}, y_{0,s}^{i,i+1} ... x_{r-1,s}^{i,i+1}, y_{r-1,s}^{i,i+1}], s \in [1, scale] \tag{13}$$

$$x_{k,s}^{i,i+1} = x_k^{i,j} + \frac{x_k^{i+1,q} - x_k^{i,j}}{scale + 1} \times s, \quad k \in \Omega \tag{14}$$

$$y_{k,s}^{i,i+1} = y_k^{i,j} + \frac{y_k^{i+1,q} - y_k^{i,j}}{scale + 1} \times s, k \in \Omega \tag{15}$$

$$x_{k,s}^{i,i+1} = 0, \quad y_{k,s}^{i,i+1} = 0, \quad k \notin \Omega \tag{16}$$

$$\Omega \in \{k \left| \left(x_k^{i,j}, y_k^{i,j}\right) \neq (0,0) \wedge \left(x_k^{i+1,q}, y_k^{i+1,q}\right) \neq (0,0)\right\} \tag{17}$$

Where $\overline{P_s^{i,i+1}}$ is the s_{th} inserted skeleton between the i_{th} and $i+1_{th}$ frame image. Ω is the set of valid key points. If a key point is not a valid key point, then the coordinates of the same key point in each inserted skeleton are set $(0,0)$. It is equivalent to regard these points as undetected points.

scale is adaptive to the specific activity. Assuming the initial number of the skeleton sequence is t, and the final number of skeleton sequence of the training sample is a fixed number denoted as T. Then the scale is formulated as:

$$\begin{aligned} &\textbf{scale} = 0, \quad t \geq 2T \\ &\textbf{scale} = \tfrac{T}{t} + 1, \quad t < 2T \end{aligned} \tag{18}$$

When the process of inserting skeletons is done, and the next step is to extract a fixed number of skeletons using a sliding window. If the amount of the skeleton sequence after inserting skeletons is $t^{'}$, the width of the sliding window is formulated as

$$\omega = \frac{t^{'}}{T} \tag{19}$$

where ω is the width of the sliding window. In each sliding window, one skeleton will be chosen randomly. An activity sample is composed of these extracted T skeletons. In theory, we can generate ω^T or even more activity samples from a raw skeleton sequence. Furthermore, all the activity samples generated from the same skeleton sequence can keep the completeness of the corresponding activity. As shown in Fig.4, the right figure contains three extracted samples from the inserted skeleton sequence. Although the space positions of the three samples are not the same, the three samples are still the same activity(lifting the left hand).

C. Human Activity Encoding

After data augmentation, all the generated activity samples is composed of T skeletons. Since there are no limitations about any factors(e.g. the type and position of the camera, the target object, the spatial position of the target object), the coordinates of the key points of the training samples are not in the same range and coordinate system.

Most current work [28] [45] solved this two problem by two strategies. First, the coordinates of the key points are transformed to the coordinates of the world coordinate system. Then, a target object will be regarded as a reference object. The coordinates of all the activity samples will be scaled to the same range. After the regularization process, the feature extraction is performed in the world coordinate system.

In fact, the three processes can be combined into one task, which is called activity encoding. In this paper, we proposed an independent activity encoding algorithm which is irrelevant to the specific dataset. Skeleton is the static posture of the target object in each frame image. In [28], cluster algorithm is applied to extract a series of the most common postures from all kinds of activities. Then the author finds out a fixed number of the most relevant postures for each type of activity using support vector machine. At last, each type of activity is encoded as a fixed number of static postures.

The encoding method is a discrete encoding method. However, the number of postures is fixed, which is not reasonable. Complex activities may have more common postures compared with simple activities. Additionally, even though different persons perform the same activity, their common postures are probably not the same exactly.

In our method, we keep all the T skeletons of each activity. Each activity consists of a series of activities. Each action is composed of the movement of all the key points. Thus, our encoding method focuses on the movement of each key point between two continuous frame images.

$$\overline{P_j^i} = [x_0^{i,j}, y_0^{i,j} \dots x_{r-1}^{i,j}, y_{r-1}^{i,j}], i \in [2, t] \tag{20}$$

$$\overline{P_q^{i-1}} = [x_0^{i-1,q}, y_0^{i-1,q} \dots x_{r-1}^{i-1,q}, y_{r-1}^{i-1,q}] \tag{21}$$

$\overline{P_j^i}$ and $\overline{P_q^{i-1}}$ are the skeleton in the i_{th} and $i-1_{th}$ frame image respectively. We focus on the joints distance between the i_{th} and $i-1_{th}$ frame image. Assuming there are **r** key points of a skeleton, then there are $r \times r$ joints distances in total. As shown in Fig.5, the blue line stands for the joints distance between the i_{th} and $i-1_{th}$ frame image. The joints distance is Euclidean distance. Then a skeleton in the i_{th} frame image is transformed into a $1 \times r^2$ feature vector $d^{i,i-1}$.

$$d_{k_1,k_2}^{i,i-1} = \sqrt{(x_{k_1}^{i,j} - x_{k_2}^{i-1,q})^2 + (y_{k_1}^{i,j} - y_{k_2}^{i-1,q})^2} \tag{22}$$

$$\overline{d^{i,i-1}} = [d_{0,0}^{i,i-1}, d_{0,1}^{i,i-1} \dots d_{r-1,r-1}^{i,i-1}] \tag{23}$$

$$i \in [2, t]; k_1 \in \Omega; k_2 \in \Omega; \tag{24}$$

where Ω is the set of valid key points between the i_{th} and $i-1_{th}$ frame image. If a key point is invalid(undetected), all the joints distances which are relevant to this key point is set $d_{\max}$. As shown in Fig.5, the red lines indicate the joints distance is invalid. $\overline{d^{i,i-1}}$ is the extracted feature vector of the skeleton in the i_{th} frame image. Since each element of the feature vector stands for the joints distance change, the feature vector of the skeleton in the 1_{th} frame image is set $\mathbf{0}$.

The skeleton sequence of an activity is composed of T feature vectors.

$$activity = \left[\overline{d^{1,0}}, \overline{d^{2,1}} \dots \overline{d^{T,T-1}}\right]^T \tag{25}$$

Except for the movement of key points, all the elements of the feature vectors are still relevant to the specific target object now. The height of the target object still affects the elements.

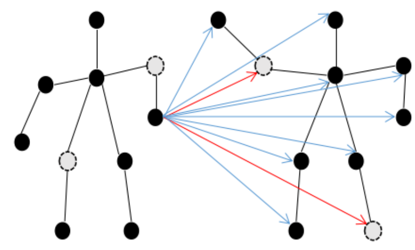


Fig. 5. (a) shows the skeleton in frame i. (b) shows the skeleton in frame i-1. The red lines indicate that it's an invalid feature.

In order to make the feature vector independent with the specific target object, all the elements of the feature vectors need to be rescaled according to the activity itself.

Assuming **max** is the maximum value of the elements except for the invalid joints distance and **min** is the minimum value of the elements, then the rescaled element is formulated as

$$g_{k_1,k_2}^{i,i-1} = 255, d_{k_1,k_2}^{i,i-1} = d_{\max} \tag{26}$$

$$g_{k_1,k_2}^{i,i-1} = \frac{d_{k_1,k_2}^{i,i-1} - \mathbf{min}}{\mathbf{max} - \mathbf{min}} \times 255, d_{k_1,k_2}^{i,i-1} \neq d_{\max} \tag{27}$$

where $g_{k_1,k_2}^{i,i-1}$ is the rescaled element of the feature vectors. In fact, it can be also regarded as the gray level of the corresponding joints distance.

Thus, the skeleton sequence of an activity sample is transformed into a $T \times r^2$ matrix:

$$activity = \left[\overline{g^{1,0}}, \overline{g^{2,1}} \dots \overline{g^{T,T-1}}\right]^T \tag{28}$$

This matrix is called activity gray matrix in this paper. It's similar to other types of gray image. Since the activity sample is transformed into activity gray matrix, the activity recognition problem is transformed into an image classification problem.

The convolutional neural network has shown outstanding performance on classification problem. However, these convolutional neural networks are trained on the datasets of color images. Thus, we are inspired to further convert activity gray matrix to color image, which can make us share the basic edge features of natural color images.

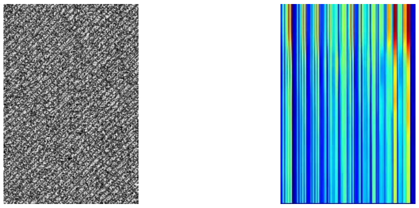
The jet colormap [46] ranging from blue to red is utilized in our method.. In order to be consistent with the range of the activity gray matrix, the jet colormap is also equally divided into 256 levels. Each level of the jet colormap contains three values of R, G, B respectively. Then the activity gray matrix is converted to a color image.

$$activity = \left[\overline{C^{1,0}}, \overline{C^{2,1}} \dots \overline{C^{T,T-1}}\right]^T \tag{29}$$

$$\overline{C^{i,i-1}} = [c_{0,0}^{i,i-1}, c_{0,1}^{i,i-1} \dots c_{r-1,r-1}^{i,i-1}] \tag{30}$$

$$c_{k_1,k_2}^{i,i-1} = jet[d_{k_1,k_2}^{i,i-1}]^T \tag{31}$$

where $jet[d_{k_1,k_2}^{i,i-1}]^T$ is a column vector which stands for the color of the corresponding joints distance $d_{k_1,k_2}^{i,i-1}$. Thus, the



(a) activity gray image (b) activity color image

Fig. 6. (a) is the activity-gray-image. (b) is the activity-color-image. (b) is converted from (a)

activity information is converted to the texture information of the color image. Fig.6 is a comparison between the activity gray matrix and the activity color image. These two images are both from the same training sample.

The color change of each column reveals the changing process of the corresponding joint distance during the activity. It's more convenient to utilize the trained model on natural color images. As for the activity gray matrix, there is no obvious texture information or the information is too confusing. Furthermore, there are no mature trained model based on large scale gray images, which makes it unable utilize shared gray texture information from other models. In our experiments, we will also evaluate the difference between training on gray matrix and training on color image.

D. Network Training

The original AlexNet was a neural network with five convolutional layers and three full-connected layers. It showed good performance on image classification. In this paper, we modify the last three full-connected layers of the AlexNet network to fit into specific human activity dataset. To avoid over-fitting , dropout is applied to the first and second full-connected layer respectively. The weights of the five convolutional layers are extracted from the pre-trained model on ILSVRC-2012.

The dropout parameter is 0.5. The number of batch size is 128. The initial learning rate is 0.0005. The Adaptive Moment Estimation(Adam) is used as the optimization algorithm. When the model is trained on the activity gray images, all the layers will be trained from scratch. When the model is trained on the activity color images, the model is fine-tuned on some layers. If the activity dataset contains **H** types of activity, the number of the output neurons is **H+1**. Assuming the output of each output neuron is O_i. The finale label of the input activity is decided as

$$label = \arg\max_i \{O_i\}, i \in [0, H] \tag{32}$$

IV. Experimental Results

In this section, we performed experiments on two public activity datasets(KARD dataset and Florence 3D dataset) and the newly released dataset(Noisy activity dataset). There are three parts of our experiments. Firstly, since the two public datasets both provide all the key points' coordinates of the target object in each frame image, we can evaluate the performance of our **Interframe Matching Algorithm** on these two datasets. Secondly, the recognition accuracy will be evaluated on these three datasets. Thirdly, the activity color images are converted from skeleton sequences. These color images are different from natural images. We will evaluate the performance of the basic texture features learned from natural images for activity color images.

A. Experiments On ***Interframe Matching Algorithm***

There are multiple persons in some activity videos of the two public datasets(KARD and Florence 3D dataset). The two datasets only provide the skeleton sequence of the target object. In our experiments, the skeleton sequence is regarded as label. We will give the skeleton a label denoted as $'1'$, which indicates that the skeleton belongs to the target object.

Then the posture estimation algorithm [40] is applied to these two public datasets, which can estimate all the potential skeletons frame by frame. Then our **Interframe Matching Algorithm** is applied to extract the skeleton sequence of the target object from the noisy skeleton sequence. At last, the extracted skeleton sequence is compared with the provided skeleton sequence frame by frame. If the two skeleton both belong to the same person, then the extraction is valid in the current frame image. The accuracy is regarded as the evaluation metric. The accuracy is formulated as

$$accuracy = \frac{N_{valid}}{N_{total}} \times 100\% \tag{33}$$

N_{valid} is the number of valid extraction while N_{total} is the total number of frame images. Tab. 1 shows the results of our **Interframe Matching Algorithm** evaluated on the two public datasets.

TABLE I
THE ACCURACY ON KARD AND FLORENCE 3D DATASET

	KARD	FLORENCE 3D
accuracy	100%	100%

Tab. 1 shows that our **Interframe Matching Algorithm** has promising performance on extracting the skeleton sequence of the target object.

B. Experiments On KARD Datasets

Kard dataset contains 18 types of activities(10 gestures and 8 activities). Each activity is performed three times by 10 persons. The dataset is divided into train dataset, validation dataset, test dataset respectively. In order to evaluate the proposed activity encoding method, the modified alexnet model is trained on activity gray matrix and converted color image respectively. Furthermore, the proposed method is compared with Gaglio [28] which is based on the provided skeleton data. Tab. 2 shows the results.

The method proposed by Gaglio [28] was based on complete skeleton data provided by skeleton tracing sensor. Our method is based on RGB images. As shown in Tab. 2, the accuracy of our model trained on the activity gray images is lower than Gaglio [28]. But our model trained on the activity color images achieves 100% accuracy.

TABLE II
FINAL ACCURACY ON KARD DATASET

	source data	Gestures	Actions
Gaglio [28]	skeleton	94.2%	96%
Our model(gray image)	RGB images	**77.5%**	**85.8%**
Our model(color image)	RGB images	**100%**	**100%**

The method proposed by Gaglio [28] is based on HMM. As we have introduced, HMM is also a timing model which attaches more importance on temporal information. Especially, the activity encoding model is a discrete model composed of a fixed amount of the most common postures. It's equivalent to recognize activities only based on a portion of temporal information. These selected pieces of temporal information are more related to activities. Thus, the method proposed by Gaglio [28] could get a good but not perfect performance. However, some temporal information and spatial information were still neglected.

Compared with Gaglio [28], our activity encoding method keeps all the spatial-temporal information. This is an adventurous idea. Its advantage is that all the related information can be kept. The disadvantage is that it's more difficult to learn the distinguishing information. Thus, our model may achieve perfect performance if the model can learn the optimal combination of all the related information. If the model doesn't catch all the related information or get the optimal combination. the accuracy may be not perfect.

Compared with the model trained on color images, the gray images based model is trained from scratch. It's equivalent to randomly search a good mapping function from the whole unknown solution space. As for the color image, activity information is transformed into color texture information. The color images based model is fine-tuned on the last several layers. It is equivalent to recursively search an optimal mapping function based on a good original state. Thus, when trained on activity color images, our model can learn the optimal combination of all the related information more easily. When the amount of training samples is fixed, it's easier to get a better performance when the model is trained on color images.

C. Experiments On Florence 3D Actions Dataset

Florence 3D dataset is also a public dataset captured by Kinect depth camera. The optional information includes RGB-D images and complete skeleton data. Here we choose several methods [47] [48] [49] [43]to compare with our method. These methods are based on RGB-D images or provided skeleton data, while our method is still only based on RGB images.

In order to be consistent, all the process on the dataset is the same with experiments on KARD dataset. Additionally, all the training parameters of convolutional neural network keep unchanging. The results of our method compared with other methods are shown in Tab. 3.

As shown in Tab. 3, we can also get the same conclusions with experimental results on KARD dataset. The model trained on activity color images still achieves 100% accuracy. However, the model trained on activity gray images still faces with the same weakness of learning the distinguishing features from scratch.

TABLE III
FINAL ACCURACY ON FLORENCE 3D DATASET

	source data	accuracy
seidenari [43]	skeleton	82%
anirudh [49]	RGB-D	89.7%
vemulapalli [48]	skeleton	90.9%
taha [47]	skeleton	96.2%
Our(gray image)	RGB images	**89.4%**
Our(color image)	RGB images	**100%**

Compared with our methods, other methods are based on RGB-D images or complete skeleton data. But our methods performed better than these methods. This is due to several factors. Firstly, depth map played a critical role in their methods. But depth map captured by depth camera was destroyed by strong noisy. Additionally, the resolution of depth map is very low. Thus, the recognition accuracy is badly influenced by the quality of depth map. Secondly, some information is neglected. Their ideas aimed at choosing the most related spatial information or temporal information. Then the recognition system recognized activities based on those selected features. The process of choosing the most related features neglected some important information.

Florence 3D actions dataset is different from KARD dataset. The duration of each activity is much smaller than activities of KARD dataset. Thus, the experimental results also prove that our method is valid regardless of the duration of activities.

*D. Experiments On **Noisy Activity Dataset***

The two public activity datasets above are very ideal. The body of the target object in each frame image is visible. Additionally, there are no obstructions and non-target objects. During the process of activity, the target object is also always faced with the camera. In order to further validate the performance of our method, the experiments are also performed on the new released **Noisy Activity Dataset**.

Each activity is performed under different conditions(no noise/ideal, obstruction, obstruction and non-target objects). All the samples which are ideal or only contain obstruction are divided into train dataset and validation dataset evenly. All the activities which contain both obstruction and non-target objects are used as test dataset in our experiment. Furthermore, we also test the performance of our method on two types of training samples(activity gray images, activity color images) respectively. The other methods are not based on incomplete skeleton sequence, so we only test the performance of our method. Tab. 4 and Tab. 5 show the detail accuracy of the model trained on gray images and color images respectively.

As shown in Tab. 4, the performance of the model trained on activity gray images is not satisfactory. Except for the same factors above, there are some other new factors. Since there are obstructions and non-target objects in our dataset, the body of the target object may be not visible exactly. Especially, if the active joints are not visible, it is more difficult to recognize the type of the activity. Thus, those methods based on RGB-D

TABLE IV
THE ACCURACY OF MODEL TRAINED ON GRAY IMAGES

	Our(gray images)
Bow	78%
Walking	92.5%
Playing tennis	53%
bending and picking up items	79%
taking shoulder bag	64%
walking while pushing chair	82.5%
bending and tying shoelace	81%

TABLE V
THE ACCURACY OF MODEL TRAINED ON COLOR IMAGES

	Our(Color images)
Bow	100%
Walking	100%
Playing tennis	100%
bending and picking up items	100%
taking shoulder bag	100%
walking while pushing chair	100%
bending and tying shoelace	100%

or complete skeleton data are not applicable any more. It's not possible to find the most common postures of the same activity. The reason can be seen that there are obstructions. As for those RGB-D images based methods, it's more difficult to detect the target object.

As shown in Tab. 5, the model trained on activity color images still achieves 100% accuracy. This is an encouraging result. The experimental results show that our method is robust to the complex realistic environment. In fact, it also proves that keeping all the spatial-temporal information is a wise choice. Although there are noises(obstructions, non-target objects interference), the convolutional neural network can still search all the available and relevant information for recognition. This is why our method achieves 100% accuracy.

Thus, based on experimental results on the two public datasets and the new released **Noisy Activity Dataset**, it can be concluded:1)Only based on the RGB images, Our method still has better performance than other methods which are based on RGB-D or ideal skeleton data. 2)Our activity encoding method based on the joints distance of two continuous frame skeleton is valid for activity recognition. 3)Natural color image converted from activity gray image can still utilize the texture feature extracted from natural color images. It can improve the performance of the model significantly.

E. Further Validation Of Texture Features Sharing

Activity color image is different from natural images. The alexnet network is composed of five convolution layers. The first layer could learn the basic texture features. When the layer is deeper, the features learned by the layer are more abstract and more relevant to the specific dataset. The experimental results above show that our model achieves perfect performance when the model is trained on activity color images. However, all the models above are trained from the first fully-connected layer. The performance of the model trained from the other layer is uncertain.

TABLE VI
THE EVALUATION RESULTS ON THREE DATASETS

	KARD	Florence 3D	**NAD**
c1-fc3	(4, 3, 758)	(1, 1, 98)	(2, 2, 89)
c2-fc3	(4, 3, 623)	(1, 1, 78)	(2, 2, 81)
c3-fc3	(3, 3, 512)	(1, 1, 66)	(2, 2, 69)
c4-fc3	(3, 3, 489)	(1, 1, 54)	(2, 2, 59)
c5-fc3	(3, 3, 89)	(1, 1, 43)	(2, 2, 41)
fc1-fc3	(3, 3, 457)	(1, 1, 55)	(2, 2, 57)

Thus, experiments are further performed on three datasets. Additionally, the conditions of experiments are set as follow:(1)the model is trained from the first convolution layer(c1-fc3). (2)the model is trained from the second convolution layer(c2-f3). (3)the model is trained from the third convolution layer(c3-fc3). (4)the model is trained from the fourth convolution layer(c4-fc3). (5)the model is trained from the fifth convolution layer(c5-fc3). (6)the model is trained from the first fully connection layer(fc1-fc3). The evaluation metrics are divided into two parts. If the accuracy of test dataset achieves 100% when the model has trained for **i** epochs on the training dataset, the first part of the evaluation metric is **i**. If the accuracy of train dataset achieves 100% for the first on the j_{th} batch data of the k_{th} epoch, the second part of the evaluation metric is (j, k). Thus, the complete result is denoted as (i, k, j).

As shown in Tab.6, it's better to train from the fifth convolution layer. Especially, it's slower to get mature model when the model is trained from the first four convolution layers or the first fully connected layer. Thus, we can know that the activity color image can utilize the basic texture information learned from natural images. Additionally, compared with training from the first fully connected layer, it's better to train from the fifth convolution layer. The features learned by the last convolutional layer is more relevant to the specific dataset.

V. THE DISCUSSION OF OUR METHOD

The experimental results show that our method achieves outstanding performance. Additionally, the proposed method is robust to obstructions and non-target object interference. Although we achieve impressive performance, there still exists some condition which may limit the performance of our method.

firstly, the proposed method is robust to non-target object interference since the target object is labeled frame by frame. Thus, at the beginning of activity, there should be no non-target objects. Secondly, during the activity process, if the target object disappears for a while due to obstructions, the non-target objects may be labeled as the target object. Then the activity may be classified incorrectly.

Although there are some limitations, the proposed method still possesses practical value. Firstly, most scenarios still satisfy the two conditions above. Secondly, our method is based on RGB images. But other methods can still utilize the key components of our method, including the Interframe Matching algorithm and activity encoding method. It means that our method can be merged into methods which aren't based on RGB images.

VI. Conclusion

In this paper, we mainly focus on the large-scale promotion of human activity recognition in elderly healthcare applications. Since the healthcare scenarios are different from industrial application scenarios in the context of Industry 4.0, the noises of the realistic environment should be taken into consideration. Thus, we proposed an effective activity recognition method which is robust to obstruction and non-target object interference. Skeleton sequence is generated from RGB images using a real-time pose estimation algorithm. Then the proposed Interframe Matching Algorithm is used to filter the non-target objects. Based on these small-scale skeleton sequences, the novel data augmentation method is applied to generate adequate training samples. At last, we design a novel encoding method which is robust to incomplete skeleton. The encoding method can convert skeleton sequence to spatial-temporal images which are fed into a CNN model for classification.

The proposed activity recognition method can be applied to various elderly healthcare applications including long-term and unobtrusive monitoring for elderly, remote medical, and primary care. Without too much demands on equipments and elderly, the robust activity recognition method is suitable for large-scale promotion.

Acknowledgment

This work is supported by the National Natural Science Foundation of China (NSFC) (No. 61231010), National High Technology Research and Development Program (No. 2015AA015901), JSPS KAKENHI Grant Number JP16K00117, JP15K15976, KDDI Foundation. Mianxiong Dong is the corresponding author.

References

[1] D. Gorecky, M. Schmitt, M. Loskyll, and D. Zühlke, "Human-machine-interaction in the industry 4.0 era," in *Industrial Informatics (INDIN), 2014 12th IEEE International Conference on*. IEEE, 2014, pp. 289–294.

[2] T. Maekawa, D. Nakai, K. Ohara, and Y. Namioka, "Toward practical factory activity recognition: unsupervised understanding of repetitive assembly work in a factory," in *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 2016, pp. 1088–1099.

[3] P. Lukowicz, J. A. Ward, H. Junker, M. Stäger, G. Tröster, A. Atrash, and T. Starner, "Recognizing workshop activity using body worn microphones and accelerometers," in *International conference on pervasive computing*. Springer, 2004, pp. 18–32.

[4] M. Aehnelt, E. Gutzeit, and B. Urban, "Using activity recognition for the tracking of assembly processes: Challenges and requirements," *WOAR*, vol. 2014, pp. 12–21, 2014.

[5] S. Tan and J. Yang, "Wifinger: leveraging commodity wifi for fine-grained finger gesture recognition," in *Proceedings of the 17th ACM International Symposium on Mobile Ad Hoc Networking and Computing*. ACM, 2016, pp. 201–210.

[6] M. S. Hossain, M. A. Rahman, and G. Muhammad, "Cyber–physical cloud-oriented multi-sensory smart home framework for elderly people: An energy efficiency perspective," *Journal of Parallel and Distributed Computing*, vol. 103, pp. 11–21, 2017.

[7] O. D. Lara and M. A. Labrador, "A survey on human activity recognition using wearable sensors." *IEEE Communications Surveys and Tutorials*, vol. 15, no. 3, pp. 1192–1209, 2013.

[8] P. Bramlage, H. Wittchen, D. Pittrow, W. Kirch, P. Krause, H. Lehnert, T. Unger, M. Höfler, B. Küpper, S. Dahm *et al.*, "Recognition and management of overweight and obesity in primary care in germany," *International journal of obesity*, vol. 28, no. 10, p. 1299, 2004.

[9] S. Weyer, M. Schmitt, M. Ohmer, and D. Gorecky, "Towards industry 4.0-standardization as the crucial challenge for highly modular, multi-vendor production systems," *Ifac-Papersonline*, vol. 48, no. 3, pp. 579–584, 2015.

[10] L. Atallah, B. Lo, R. King, and G.-Z. Yang, "Sensor positioning for activity recognition using wearable accelerometers," *IEEE transactions on biomedical circuits and systems*, vol. 5, no. 4, pp. 320–329, 2011.

[11] J. Shotton, T. Sharp, A. Kipman, A. Fitzgibbon, M. Finocchio, A. Blake, M. Cook, and R. Moore, "Real-time human pose recognition in parts from single depth images," *Communications of the ACM*, vol. 56, no. 1, pp. 116–124, 2013.

[12] A. F. Bobick and J. W. Davis, "The recognition of human movement using temporal templates," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 23, no. 3, pp. 257–267, 2001.

[13] A. Iosifidis, N. Nikolaidis, and I. Pitas, "Movement recognition exploiting multi-view information," in *Multimedia Signal Processing (MMSP), 2010 IEEE International Workshop on*. IEEE, 2010, pp. 427–431.

[14] A. Iosifidis, A. Tefas, and I. Pitas, "Multi-view action recognition based on action volumes, fuzzy distances and cluster discriminant analysis," *Signal Processing*, vol. 93, no. 6, pp. 1445–1457, 2013.

[15] I. Laptev, "On space-time interest points," *International journal of computer vision*, vol. 64, no. 2-3, pp. 107–123, 2005.

[16] J. Parkka, M. Ermes, P. Korpipaa, J. Mantyjarvi, J. Peltola, and I. Korhonen, "Activity classification using realistic data from wearable sensors," *IEEE Transactions on information technology in biomedicine*, vol. 10, no. 1, pp. 119–128, 2006.

[17] M. Ermes, J. Pärkkä, J. Mäntyjärvi, and I. Korhonen, "Detection of daily activities and sports with wearable sensors in controlled and uncontrolled conditions," *IEEE transactions on information technology in biomedicine*, vol. 12, no. 1, pp. 20–26, 2008.

[18] J. E. Berlin, K. L. Storti, and J. S. Brach, "Using activity monitors to measure physical activity in free-living conditions," *Physical Therapy*, vol. 86, no. 8, pp. 1137–1145, 2006.

[19] A. Nazábal, P. García-Moreno, A. Artés-Rodríguez, and Z. Ghahramani, "Human activity recognition by combining a small number of classifiers," *IEEE journal of biomedical and health informatics*, vol. 20, no. 5, pp. 1342–1351, 2016.

[20] S. Ha and S. Choi, "Convolutional neural networks for human activity recognition using multiple accelerometer and gyroscope sensors," in *Neural Networks (IJCNN), 2016 International Joint Conference on*. IEEE, 2016, pp. 381–388.

[21] A. H. Al-Fatlawi, H. K. Fatlawi, and S. H. Ling, "Recognition physical activities with optimal number of wearable sensors using data mining algorithms and deep belief network," in *Engineering in Medicine and Biology Society (EMBC), 2017 39th Annual International Conference of the IEEE*. IEEE, 2017, pp. 2871–2874.

[22] N. Hegde, M. Bries, T. Swibas, E. Melanson, and E. Sazonov, "Automatic recognition of activities of daily living utilizing insole based and wrist worn wearable sensors," *IEEE Journal of Biomedical and Health Informatics*, 2017.

[23] S. Scheurer, S. Tedesco, K. N. Brown, and B. O'Flynn, "Human activity recognition for emergency first responders via body-worn inertial sensors," in *Wearable and Implantable Body Sensor Networks (BSN), 2017 IEEE 14th International Conference on*. IEEE, 2017, pp. 5–8.

[24] T. Maekawa, Y. Kishino, Y. Yanagisawa, and Y. Sakurai, "Wristsense: wrist-worn sensor device with camera for daily activity recognition," in *Pervasive Computing and Communications Workshops (PERCOM Workshops), 2012 IEEE International Conference on*. IEEE, 2012, pp. 510–512.

[25] C. Laoudias, P. Tsangaridis, M. Polycarpou, C. Panayiotou, C. Kyrkou, and T. Theocharides, "Cooperative fault-tolerant target tracking in camera sensor networks," in *Communications (ICC), 2015 IEEE International Conference on*. IEEE, 2015, pp. 6634–6639.

[26] Z. Zhang and S. Poslad, "Improved use of foot force sensors and mobile phone gps for mobility activity recognition," *IEEE Sensors Journal*, vol. 14, no. 12, pp. 4340–4347, 2014.

[27] S. Lv, Y. Lu, M. Dong, X. Wang, Y. Dou, and W. Zhuang, "Qualitative action recognition by wireless radio signals in human–machine systems," *IEEE Transactions on Human-Machine Systems*, vol. 47, no. 6, pp. 789–800, 2017.

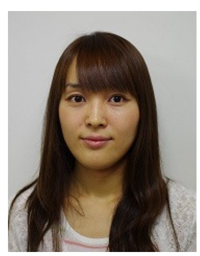
[28] S. Gaglio, G. L. Re, and M. Morana, "Human activity recognition process using 3-d posture data," *IEEE Transactions on Human-Machine Systems*, vol. 45, no. 5, pp. 586–597, 2015.

[29] E. Cippitelli, E. Gambi, S. Spinsante, and F. Flórez-Revuelta, "Evaluation of a skeleton-based method for human activity recognition on a large-scale rgb-d dataset," 2016.

[30] D. Wu, L. Pigou, P.-J. Kindermans, N. D.-H. Le, L. Shao, J. Dambre, and J.-M. Odobez, "Deep dynamic neural networks for multimodal gesture segmentation and recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 8, pp. 1583–1597, 2016.
[31] J. Wan, G. Guo, and S. Z. Li, "Explore efficient local features from rgb-d data for one-shot learning gesture recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 8, pp. 1626–1639, 2016.
[32] Y. Shin and K.-J. Yoon, "Patchmatch belief propagation meets depth upsampling for high-resolution depth maps," *Electronics Letters*, vol. 52, no. 17, pp. 1445–1447, 2016.
[33] W. Liu, X. Chen, J. Yang, and Q. Wu, "Robust color guided depth map restoration," *IEEE Transactions on Image Processing*, vol. 26, no. 1, pp. 315–327, 2017.
[34] P. Sensor, "Nite 1.3 algorithms notes, version 1.0, primesense inc. 2010."
[35] B. Mahasseni and S. Todorovic, "Regularizing long short term memory with 3d human-skeleton sequences for action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3054–3062.
[36] Y. Hou, Z. Li, P. Wang, and W. Li, "Skeleton optical spectra based action recognition using convolutional neural networks," *IEEE Transactions on Circuits and Systems for Video Technology*, 2016.
[37] Y. Du, W. Wang, and L. Wang, "Hierarchical recurrent neural network for skeleton based action recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1110–1118.
[38] G. Ning, Z. Zhang, and Z. He, "Knowledge-guided deep fractal neural networks for human pose estimation," *arXiv preprint arXiv:1705.02407*, 2017.
[39] Q. Fu, Q. Quan, and K.-Y. Cai, "Robust pose estimation for multirotor uavs using off-board monocular vision," *IEEE Transactions on Industrial Electronics*, 2017.
[40] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," *arXiv preprint arXiv:1611.08050*, 2016.
[41] M. Chen, W. Lin, and B. Zhou, "A real-time virtual dressing system with rgb-d camera," in *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2015 Asia-Pacific*. IEEE, 2015, pp. 1041–1044.
[42] X. Liu, G.-F. He, S.-J. Peng, Y.-m. Cheung, and Y. Y. Tang, "Efficient human motion retrieval via temporal adjacent bag of words and discriminative neighborhood preserving dictionary learning," *IEEE Transactions on Human-Machine Systems*, 2017.
[43] L. Seidenari, V. Varano, S. Berretti, A. Bimbo, and P. Pala, "Recognizing actions from depth cameras as weakly aligned multi-part bag-of-poses," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2013, pp. 479–485.
[44] C. Li, P. Wang, S. Wang, Y. Hou, and W. Li, "Skeleton-based action recognition using lstm and cnn," in *Multimedia & Expo Workshops (ICMEW), 2017 IEEE International Conference on*. IEEE, 2017, pp. 585–590.
[45] C.-H. Kuo, P.-C. Chang, and S.-W. Sun, "Behavior recognition using multiple depth cameras based on a time-variant skeleton vector projection," *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2017.
[46] P. Wang, Z. Li, Y. Hou, and W. Li, "Action recognition based on joint trajectory maps using convolutional neural networks," in *Proceedings of the 2016 ACM on Multimedia Conference*. ACM, 2016, pp. 102–106.
[47] A. Taha, H. H. Zayed, M. Khalifa, and E.-S. M. El-Horbaty, "Human activity recognition for surveillance applications," in *Proceedings of the 7th International Conference on Information Technology*, 2015, pp. 577–586.
[48] R. Vemulapalli, F. Arrate, and R. Chellappa, "Human action recognition by representing 3d skeletons as points in a lie group," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 588–595.
[49] R. Anirudh, P. Turaga, J. Su, and A. Srivastava, "Elastic functional coding of human actions: From vector-fields to latent variables," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3147–3155.

Yi Chen received his B.E. degree from the Department of Electronics and Information Engineering, Huazhong University of Science and Technology(HUST), Wuhan, China, in 2016. Currently, he is pursuing his M.S. degree in the School of Electronic Information and Communications, HUST. From Aug. 2017, he is also a Visiting Scholar in the Emerging Networks and Systems Lab and Wireless Networks Lab, the Department of Information and Electronic Engineering, Muroran Institute of Technology, Muroran, Hokkaido, Japan, under the guidance of Prof. Kaoru Ota. His current research interests include computer vision, artificial intelligence.

Li Yu received her B.Sc. and Ph.D. degrees from the Department of Electronics and Information Engineering, Huazhong University of Science and Technology, Wuhan, China, in 1992 and 1999, respectively. She joined Huazhong University of Science and Technology, in 1999, where she is now a Professor at the School of Electronic Information and Communications. Her current research interests include image and video coding, multimedia communications, social networks and wireless networking.



Kaoru Ota was born in Aizu-Wakamatsu, Japan. She received M.S. degree in Computer Science from Oklahoma State University, USA in 2008, B.S. and Ph.D. degrees in Computer Science and Engineering from The University of Aizu, Japan in 2006, 2012, respectively. She is currently an Assistant Professor with Department of Information and Electronic Engineering, Muroran Institute of Technology, Japan. From March 2010 to March 2011, she was a visiting scholar at University of Waterloo, Canada. Also she was a Japan Society of the Promotion of Science (JSPS) research fellow with Kato-Nishiyama Lab at Graduate School of Information Sciences at Tohoku University, Japan from April 2012 to April 2013. Her research interests include Wireless Networks, Cloud Computing, and Cyber-physical Systems. Dr. Ota has received best paper awards from ICA3PP 2014, GPC 2015, IEEE DASC 2015, IEEE VTC 2016-Fall, FCST 2017 and IET Communications 2017. She is an editor of IEEE Transactions on Vehicular Technology (TVT), IEEE Communications Letters, Peer-to-Peer Networking and Applications (Springer), Ad Hoc & Sensor Wireless Networks, International Journal of Embedded Systems (Inderscience) and Smart Technologies for Emergency Response & Disaster Management (IGI Global), as well as a guest editor of ACM Transactions on Multimedia Computing, Communications and Applications (leading), IEEE Communications Magazine, IEEE Network, etc. Also she was a guest editor of IEEE Wireless Communications (2015), IEICE Transactions on Information and Systems (2014), and Ad Hoc & Sensor Wireless Networks (Old City Publishing) (2014). She was a research scientist with A3 Foresight Program (2011-2016) funded by Japan Society for the Promotion of Sciences (JSPS), NSFC of China, and NRF of Korea. She is the recipient of IEEE TCSC Early Career Award 2017.

Mianxiong Dong received B.S., M.S. and Ph.D. in Computer Science and Engineering from The University of Aizu, Japan. He is currently an Associate Professor in the Department of Information and Electronic Engineering at the Muroran Institute of Technology, Japan. He was a JSPS Research Fellow with School of Computer Science and Engineering, The University of Aizu, Japan and was a visiting scholar with BBCR group at University of Waterloo, Canada supported by JSPS Excellent Young Researcher Overseas Visit Program from April 2010 to August 2011. Dr. Dong was selected as a Foreigner Research Fellow (a total of 3 recipients all over Japan) by NEC C&C Foundation in 2011. His research interests include Wireless Networks, Cloud Computing, and Cyber-physical Systems. He has received best paper awards from IEEE HPCC 2008, IEEE ICESS 2008, ICA3PP 2014, GPC 2015, IEEE DASC 2015, IEEE VTC 2016-Fall, FCST 2017 and 2017 IET Communications Premium Award. Dr. Dong serves as an Editor for IEEE Transactions on Green Communications and Networking (TGCN), IEEE Communications Surveys and Tutorials, IEEE Network, IEEE Wireless Communications Letters, IEEE Cloud Computing, IEEE Access, as well as a leading guest editor for ACM Transactions on Multimedia Computing, Communications and Applications (TOMM), IEEE Transactions on Emerging Topics in Computing (TETC), IEEE Transactions on Computational Social Systems (TCSS). He has been serving as the Vice Chair of IEEE Communications Society Asia/Pacific Region Meetings and Conference Committee, Leading Symposium Chair of IEEE ICC 2019, Student Travel Grants Chair of IEEE GLOBECOM 2019, and Symposium Chair of IEEE GLOBECOM 2016, 2017. He is the recipient of IEEE TCSC Early Career Award 2016, IEEE SCSTC Outstanding Young Researcher Award 2017 and The 12th IEEE ComSoc Asia-Pacific Young Researcher Award 2017.