Pattern Recognition 98 (2020) 107053

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

Multi-stage adaptive regression for online activity recognition

Bangli Liu[a], Haibin Cai[b], Zhaojie Ju[a], Honghai Liu[a,*]

[a] Intelligent Systems and Biomedical Robotics Group, School of Computing, University of Portsmouth, UK
[b] Computer Science, Loughborough University, UK

ARTICLE INFO

Article history:
Received 8 October 2018
Revised 16 June 2019
Accepted 12 September 2019
Available online 13 September 2019

Keywords:
Online activity recognition
Interaction recognition
Partial observation
Adaptive regression

ABSTRACT

Online activity recognition which aims to detect and recognize activity instantly from a continuous video stream is a key technology in human-robot interaction. However, the partial activity observation problem, mainly due to the incomplete sequence acquisition, makes it greatly challenging. This paper proposes a novel approach, named Multi-stage Adaptive Regression (MAR), for online activity recognition with the main focus on addressing the partial observation problem. Specifically, the MAR framework delicately assembles overlapped activity observations to improve its robustness against arbitrary activity segments. Then multiple score functions corresponding to each specific performance stage are collaboratively learned via a adaptive label strategy to enhance its power of discriminating similar partial activities. Moreover, the Online Human Interaction (OHI) database is constructed to evaluate the online activity recognition in human interaction scenarios. Extensive experimental evaluations on the Multi-Modal Action Detection (MAD) database and the OHI database show that the MAR method achieves an outstanding performance over the state-of-the-art approaches.

1. Introduction

Human activity recognition has received increasing attention in computer vision due to its wide applications in human behavior analysis, public surveillance, healthcare assistance, human-robot interaction, and smart homes [1–3]. Human motion data can be acquired via RGB cameras, RGBD sensors or wearable sensors. Most of the early research uses only the color and texture information in 2D images, thus their recognition performance might be limited due to the lack of 3D information. Although some wearable sensor-based motion capture systems [4] have been proposed to obtain 3D skeleton data, the inconvenience of wearing extra devices hinders it from being applied to the general public. Fortunately, the emergence of the cost-effective RGBD sensors has largely alleviated these shortcomings by providing depth images and 3D skeleton data [5]. However, it still remains a challenge for online activity recognition in practical scenarios.

Traditional offline activity recognition approaches aim to recognize pre-segmented activities whose start time and end time are manually extracted [6–14]. However, in most practical scenarios, it is difficult to know the boundary of activities ahead and the recognition results need to be given during the activity period with low latency. For example, an alarm is expected to be triggered immediately if a dangerous behavior is going to happen in a public surveillance scenario, and an assisted robot should be able to identify the danger and provide timely help before elderly people fall down. Thus, online activity recognition which aims to detect and recognize activities as soon as possible in a continuous video stream is extremely important in such applications.

Recent research has tried to recognize an ongoing activity from the observation in its early stage [15–20], namely early activity recognition, which solves the problem to some extent. For example, Huang et al. [19] made use of the bag-of-words to depict features of body joints and built a sequential max-margin event detector (SMMED) for early activity recognition. They assumed that one of given actions should be happening in each testing video stream, and the final action label was identified by discarding the actions with decreasing probability. Cai et al. [20] developed a low latency system to identify actions when enough observations are obtained. The most likely action class is selected depending on the vote of each action within a temporal window. Despite activities can be recognized before they are fully completed, the requirement of knowing the content or starting time of an activity ahead greatly limits their application in practical scenarios.

Compared to offline activity recognition and early activity recognition, online activity recognition is more complex in that it needs to simultaneously and quickly perform activity detection and recognition in a continuous video stream without any prior knowledge of the action start time and end time [21], as shown in Fig. 1.

* Corresponding author.
E-mail address: honghai.liu@port.ac.uk (H. Liu).

https://doi.org/10.1016/j.patcog.2019.107053
0031-3203/

Fig. 1. Offline activity recognition aims to classify fully observed activities that are pre-segmented according to activity categories, the goal of early activity recognition is to recognize an activity with the partially observed activity from the beginning, and online activity recognition intends to detect and recognize activity as soon as possible from a continuous video stream without any manual segmentation. The partial activity observation is a common but challenging problem in online activity recognition.

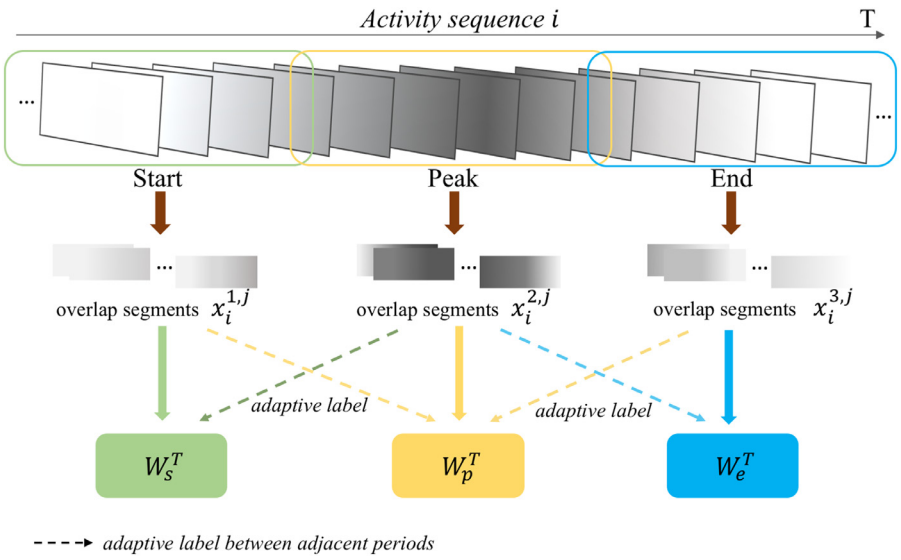


Fig. 2. The overview of the proposed MAR framework. $\boldsymbol{W}_s^T, \boldsymbol{W}_p^T, \boldsymbol{W}_e^T$ are score functions corresponding to each performance stage, which are collaboratively learned by adaptive labeling segments from adjacent performance stages. The color gradient indicates the general trend of the information of an activity along the time domain. Segments in the dark color contain more information than segments in the light color, and the information becomes more while the color becoming darker.

One of the most challenging problems in online activity recognition is the partial activity observation problem that only part of the activity can be observed due to the incomplete sequence acquisition. The partial activity observation is possible at arbitrary performance stages, with a large inter-class variability and intra-class similarity, and the available information of activities is limited.

This paper proposes a novel Multi-stage Adaptive Regression (MAR) framework to address the partial activity observation problem in online activity recognition. Fig. 2 shows the structure of MAR framework, where human activities are roughly divided into three performance stages and segments from adjacent stages are used via an adaptive label strategy to remain the consistency of sequential data. Different activities may have very similar observations at their starting or ending stages and uniformly using these observations for training might result in poor recognition performance. The discriminative power of the regression model is enhanced by collaboratively learning three score functions corresponding to three different stages. Considering that it is difficult to decide the exact boundary between performance stages of activities, the inherent evolution of segments from adjacent performance stages is modeled by introducing an adaptive label strategy into the learning formulation. Note that the focus of this paper is on exploring the feasibility of solving partial activities via multi-stage adaptive regression rather than accurately identifying the exact three action stages. To partition the three stages, a manual labeling process is required in the training dataset. The proposed method is capable of simultaneously detecting and recognizing activities from untrimmed videos. The main contributions of this paper are summarized as follows:

1. An effective multi-stage adaptive regression framework is proposed to address the partial activity observation problem in online activity recognition.
2. Multiple score functions are collaboratively learned via an adaptive label strategy to reinforce the capacity of distinguishing similar partial activities and the robustness to arbitrary activity fragments.
3. A challenging interaction database, Online Human Interaction (OHI), is collected in a realistic scenario to further evaluate the online activity recognition.

The remainder of this paper is organized as follows: Section 2 reviews the related work of human activity recognition. Section 3 introduces the proposed method. Section 4 reports various experimental results as well as the comparison with the state-of-the-art methods. Section 5 summarizes the work of this paper.

2. Related work

This section reviews three related research tasks: Offline activity recognition, early activity recognition, and online activity recognition. Fig. 1 gives an overview of the differences among these three tasks.

2.1. Offline activity recognition

Offline activity recognition methods aim to classify activities that are already pre-segmented by manually providing the starting and ending points of activities. Recent years have seen significant progress in 3D skeleton joints base human activity recognition [14,22–28]. Vemulapalli et al. [23] used the rotation and translation information among body parts to model their relative geometry relationships, with which human movements were encoded to curves in the Lie group. Guo et al. [25] captured the gradient-based invariants in the dual square-root function to describe the motion trajectories. Then human actions were recognized by combining the trajectories of different rigid body parts. In [26], the geometric and motion feature of joints were calculated to learn a dictionary of body poses or atomic actions, whose sparse composition were used to represent complex human activities.

The great achievements of the deep learning technique in the image classification task has motivated its increasing application in human activity recognition [29–33]. Yan et al. [29] proposed to model the dynamic information of skeleton data by connecting skeleton joints spatially and temporally in a multi-layer graph neural networks. Liu et al. [30] proposed to encode skeleton information such as location and velocity into skeleton images, from which the spatio-temporal correlations among joints were learned via a modified Inception-ResNet convolution neural networks. Recurrent neural networks (RNN) [11,34–36] have also been adopted for skeleton-based action recognition, owing to its appealing capacity of modeling temporal information of action sequences. Zhang et al. [11] proposed a multi-stream long short-term memory framework to learn different geometric features which were further combined with a smoothed score fusion approach for classification. Veeriah et al. [35] fed various hand-crafted features to a differential RNN which improves traditional LSTM by adding a gate function to model the dynamics of salient motions. Du et al. [34,37] proposed an end-to-end hierarchical RNN which fuses the feature extracted from five human body parts for action recognition. To extract the relationship between body parts, Shahroudy et al. [36] utilized the human body structure to build a part-aware LSTM. By concatenating part-based memory cells, the non-adjacent parts relations were learned from the 3D skeleton sequence.

2.2. Early activity recognition

With an assumption that the starting time of an action can be known ahead, early activity recognition targets at recognizing activities before they are completely executed and ideally at the beginning of their occurrences. In [15], the dynamic bag-of-words was developed to model the change of feature distributions of human activities as the observation percentage increasing. Then the unfinished activities were identified by a probabilistic model based on available information. Hoai et al. [18] presented an early event detection method, where a structured output SVM was trained to detect the action via a threshold strategy. Huang et al. [19] proposed a sequential max-margin event detector for early action recognition. With the increasing of observed frames, actions showing decreasing probability of the occurrence will be discarded. Escalante et al. [38] argued that the implementation of the naive Bayes was simple and fast for early gesture recognition. They used complete activity sequences to train the naive Bayes classifiers and classified a new sequence from the beginning by accumulating temporal information. Hu et al. [39] proposed to classify actions in their different progress percentages and accumulate the knowledge for a final decision. Different from the aforementioned works which require the assumption of knowing the exact start time of the actions ahead, this paper overcomes this assumption by proposing an adaptive label strategy to model the inherent evolution of activity segments from adjacent performance stages for online action recognition.

2.3. Online activity recognition

Online activity recognition aims to detect and recognize activities as soon as possible in a continuous video stream [40]. Reyes et al. [41] proposed an architecture to identify the transitions between human activities using data from inertial sensors, which was achieved by using the probabilistic results of activities from an SVM classifier and a heuristic filter. Kulkarni et al. [42] utilized dynamic time warping and dynamic frame warping for simultaneous action segmentation and classification. In their method, each video frame was assigned a label based on its comparison to the template representations, and the change of labels between consecutive frames indicated the starting or ending point of an event. Similarly, Wu et al. [43] clustered action clips into several action-words and learned an action-topics model to reflect the co-occurrence and temporal relations of the action-words. The action segmentation was realized according to the convert of action topics between consecutive clips. Instead of segmenting video streams to isolated activities, Nowozin et al. [44] proposed to detect action points which functioned as action peak frames to speed up the detection performance for online action recognition. Based on this, some approaches were developed for the action points detection in streaming videos [45–48]. Bloom et al. [48] proposed to combine the clustered spatio-temporal manifolds and the temporal history of activities to detect the peaks of actions in a continuous stream. However, the detection result of a single time instance might not be representative enough for the complete action sequences and can cause false detections especially when the peak frames from different activities are quite similar.

Recently, many deep learning methods [49–54] have been proposed for online action recognition in video streams. Liu et al. [55] proposed a large scale action recognition dataset named PKU-MMD and explored several LSTM based methods for action detection. Molchanov et al. [49] combined a recurrent three-dimensional convolutional neural network with connectionist temporal classification [56] to jointly detect and recognize gestures without requiring prior segmentation. Shou et al. [52] utilized multi-stage CNNs to generate candidate action segments for recognition. Similarly, Chai et al. [53] firstly segmented continuous gesture video streams into isolated gestures and then recognized the segmented gestures by fusing multi-modal features in a two streams RNN framework. Wang et al. [54] proposed to learn the geometric relations among joints for action recognition and detection by introducing a genetic end-to-end RNN based network. A frame-wise classification strategy and a multi-scale sliding window search algorithm were used for action detection.

2.4. Motivation

The offline activity recognition methods suffer from a drawback that the recognition results can only be given after the action event. Observing this problem, many researchers proposed the early activity recognition to obtain the recognition results during the activity period. However, it still relies on pre-segmented sequences by providing the starting point of the activity. To deal with the previous problems, online activity recognition has been appealing in recent research. The partial activity observation problem mainly caused by the incomplete sequence acquisition, makes it greatly challenging due to following reasons: 1) The partial observation is possible from an arbitrary performance stage, with a large inter-class variability and intra-class similarity; 2) activities need to be recognized as accurately as possible based on the limited information from the partially observed activities.

To mitigate this issue, this paper proposes a Multi-stage Adaptive Regression (MAR) framework, where multiple score functions are collaboratively learned corresponding to each performance stage. Activity segments spanning all performance stages are collected during the learning stage to make score functions robust and effective to arbitrary activity fragments. Furthermore, the inherent evaluation of segments from adjacent performance stages is considered by introducing an adaptive label strategy into the learning formulation. By doing this, the MAR method is capable of identifying activities from partial observations with an outperforming performance.

3. Multi-stage adaptive regression framework

3.1. Problem statement

This paper aims to develop an online activity recognition method for identifying ongoing activity sequences. The method is particularly designed to deal with the partial activity observation problem, and it is required to be robust to any activity segment. Based on the evolution of an activity along the time domain, an activity is progressively divided into three stages: *start, peak*, and *end*. The *start* segments are in an onset stage describing the transition from the initial status to the *peak* status which includes the most salient information of an activity of interest, while the *end* segments are in an offset stage depicting the transition from the *peak* status back to the *end* status. Given a fully observed activity sequence $X[1{:}\ T]$ of length T, an arbitrary partial observation of it can be represented by $X[t_1{:}\ t_2]$, where $1 \le t_1 < t_2 \le T$. Herein, t_1 is not constrained to be 1, which means that the partial activity could be at any performance stage. This overcomes the assumption that the partial activity needs to be observed from the start of an activity in [15,38]. Note, T might vary in different activity sequences. Our goal is to extract discriminative information for any partial activity $X[t_1{:}\ t_2]$ and then assign an activity label to it.

3.2. 3D spatio-temporal activity representation

Recently, a lot of human activity recognition approaches based on skeleton joints have achieved satisfactory performance, owing to the invariance of the skeleton information to different locations and human appearances [57].

In a previous work [58], we have introduced a novel spatio-temporal feature descriptor for effective offline human action recognition. In this paper, the descriptor is adapted for the online action recognition scenario. The coordinates of skeleton joints are firstly transformed to a person-centric coordinate system using the following equation:

$$C = R * C^{'} + T \tag{1}$$

where C and $C^{'}$ stand for the original and transformed coordinate, respectively, and T represents the coordinate of the body center. $R = [a_1; a_2; a_3]^{-1}$ is the rotation matrix. a_1, a_2, and a_3 are the unit vector of $x^{'}$, $y^{'}$, and $z^{'}$ axis in the proposed person-centric coordinate system, as shown in Fig. 3(a). The $x^{'}$ axis is the normal vector of a plane constructed by the spine point, the left hip point and the right hip point, the $z^{'}$ axis is calculated using the vector from the hip center to the spine joint, and the $y^{'}$ axis is determined by the dot product of the $x^{'}$ and $z^{'}$ axis. This transformation reduces the influence of various locations and orientations between subjects and the sensor.

The histogram of moving trend is mined from the frame moving directions $\mathbf{V}_t$, which can effectively infer the intention of an activity at a high level. The directions in 3D space are empirically decomposed into l semantic moving words $\bar{\mathbf{V}}_d, 1 \le d \le l$. Finally, the motion feature over a sequence is coded to the moving trend using the *cosine* similarity. For the geometry feature, it is improved to make it adaptive to the segment-wise recognition by using the start of the segment as initial status x_{in}, y_{in}, z_{in}. Therefore, the process of feature extraction is summarized as follows:

$$\begin{cases} \mathbf{v_t^i} = \{x_{p_t^i} - x_{p_{t-1}^i}, y_{p_t^i} - y_{p_{t-1}^i}, z_{p_t^i} - z_{p_{t-1}^i}\} \\ \cos\theta_d^i(t) = \frac{\mathbf{v_t^i} \cdot \bar{\mathbf{v}}_\mathbf{d}}{\|\mathbf{v_t^i}\| \|\bar{\mathbf{v}}_\mathbf{d}\|}, d \in [1, l] \\ bin_d = \sum_{t=t1}^{t2} \|\mathbf{v}_t^i\| \times max\{cos\theta_d^i(t)\}, d \in [1, l] \\ H(i) = \{bin_1, \ldots, bin_m\} \\ \triangle d_t^i = \{x_{p_t^i} - x_{in}^i, y_{p_t^i} - y_{in}^i, z_{p_t^i} - z_{in}^i\} \\ G(t) = \{\triangle d_t^1, \ldots, \triangle d_t^N\} \end{cases} \tag{2}$$

where $x_{p_t^i}, y_{p_t^i}, z_{p_t^i}$ represent the transformed coordinates of the ith joint. $\mathbf{v_t^i}$ and $\bar{\mathbf{v}}_\mathbf{j}$ denote the frame-level moving direction and the semantic moving word. bin_d indicates the dth bin in the histogram of the moving trend feature, and $\triangle d_t^i$ represents the relative displacement of the ith joint at frame t. $H(i)$ and $G(t)$ mean the extracted moving trend feature and geometry feature, respectively.

3.3. Multi-stage adaptive regression

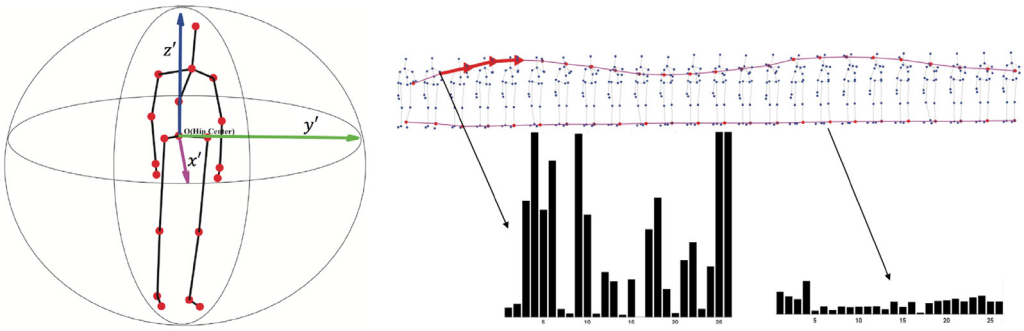
Let $(X_1, \boldsymbol{y}_1), (X_2, \boldsymbol{y}_2), \ldots, (X_n, \boldsymbol{y}_n)$ be the training data, where X_i is the ith activity sample and $\boldsymbol{y}_i$ is the label vector. To cope with random activity fragment $X[t_1{:}\ t_2]$, the segment $x_i^{k,j} (x_i^{k,j} \in X_i)$ at each performance stage k from each category i is generated for learning, as shown in Fig. 2. A preliminary idea is to learn a single score function to measure the compatibility between activity fragments and labels using all segments, as denoted by the following formula:

$$\begin{aligned} \min_{\boldsymbol{W}_o} \sum_{i=1}^{n} \sum_{k=1}^{3} \sum_{j=1}^{m_k} & ||\boldsymbol{W}_o^T \quad \Phi(x_i^{k,j}) - \boldsymbol{y}_i||_{1,2} + \frac{\xi}{2}||\boldsymbol{W}_o||_2^2 \\ & s.t. \quad \xi \ge 0. \end{aligned} \tag{3}$$

where m_k is the number of activity segments from the performance stage k, $\boldsymbol{W}_o$ is the score function, and $\Phi(\cdot)$ is the proposed feature extraction function. We refer this method as Multi-stage Regression (MR) since activity segments at multiple stages are used for the regression function. The MR method has the capacity of identifying the partial activity when it happens at the *peak* stage as it includes the most salient information of an activity of interest. However, the power of discriminating similar segments from the *start* or *end* stage is insufficient, which will be discussed in Section 4.

On top of this, an enhanced regression framework, named Multi-stage Adaptive Regression (MAR) is introduced. MAR can be regarded as a fine-grained regression framework, where multiple score functions for each specific performance stage are collaboratively learned to improve the power of discriminating similar activity segments. To remain the consistency among sequential segments, an adaptive label strategy is modeled for adjacent stages. The MAR method is formulated as follows:

$$\begin{aligned} \min_{\boldsymbol{W},\lambda} & \sum_{i=1}^{n} \sum_{k=1}^{2} \sum_{j=1}^{m_k} ||\boldsymbol{W}_s^T \Phi(x_i^{k,j}) - \boldsymbol{y}_i(1 - \lambda_{s,j}|Sgn(k-1)|)||_{1,2} \\ & + \sum_{i=1}^{n} \sum_{k=1}^{3} \sum_{j=1}^{m_k} ||\boldsymbol{W}_p^T \Phi(x_i^{k,j}) - \boldsymbol{y}_i(1 - \lambda_{p,j}|Sgn(k-2)|)||_{1,2} \\ & + \sum_{i=1}^{n} \sum_{k=2}^{3} \sum_{j=1}^{m_k} ||\boldsymbol{W}_e^T \Phi(x_i^{k,j}) - \boldsymbol{y}_i(1 - \lambda_{e,j}|Sgn(k-3)|)||_{1,2} \\ & + \frac{\xi_s}{2}||\boldsymbol{W}_s||_2^2 + \frac{\xi_p}{2}||\boldsymbol{W}_p||_2^2 + \frac{\xi_e}{2}||\boldsymbol{W}_e||_2^2 \\ & s.t. \quad \xi_s, \xi_p, \xi_e \ge 0, \quad 0 \le \lambda_{s,j}, \lambda_{p,j}, \lambda_{e,j} \le 1 \end{aligned} \tag{4}$$



(a) (b)

Fig. 3. (a) Person-centric coordinate system. (b) The moving trend of skeleton joints.

where $\boldsymbol{W}_s^T, \boldsymbol{W}_p^T, \boldsymbol{W}_e^T$ are score functions corresponding to the *start*, *peak*, and *end* stage, respectively. $Sgn(\cdot)$ is the sign function and m_k is the number of segments in each stage. $(1-\lambda_{s,j}|Sgn(k-1)|)$ is the weight of segments to emphasize the importance of segments from k stage over its adjacent stage. For example, when learning W_s^T, $(1-\lambda_{s,j}|Sgn(k-1)|)$ will be 1 if segment $\Phi(x_i^{k,j})$ is selected from *start* stage, otherwise it will be less than 1. Thus, Eq. (4) can be rewritten into simplified form as follows:

$$\begin{aligned}&\min_{\boldsymbol{W},\theta}\sum_{i=1}^{n}||\boldsymbol{W}_s^T\widehat{\mathbf{X}}_i-\boldsymbol{Y}_i\theta_{sp}||_{1,2}+\frac{\xi_s}{2}||\boldsymbol{W}_s||_2^2\\&+||\boldsymbol{W}_p^T\widehat{\mathbf{X}}_i-\boldsymbol{Y}_i\theta_{spe}||_{1,2}+\frac{\xi_p}{2}||\boldsymbol{W}_p||_2^2\\&+||\boldsymbol{W}_e^T\widehat{\mathbf{X}}_i-\boldsymbol{Y}_i\theta_{pe}||_{1,2}+\frac{\xi_e}{2}||\boldsymbol{W}_e||_2^2\\&s.t.\quad \xi_s,\xi_p,\xi_e\geq 0,\quad 0\leq\theta_{sp},\theta_{spe},\theta_{pe}\leq 1\end{aligned}\tag{5}$$

where $\widehat{\mathbf{X}}$ denotes the feature descriptor matrix of segments, and $\boldsymbol{Y}$ is the corresponding label matrix. The elements in $\boldsymbol{\theta}_{sp}$ corresponding to the segments at the *start* stage are constrained to be 1 while learning $\boldsymbol{W}_s^T$ (by analogy, $\boldsymbol{\theta}_{spe}$ and $\boldsymbol{\theta}_{pe}$ have the same affinity).

Compared to MR, there are three score functions $\boldsymbol{W}_s^T, \boldsymbol{W}_p^T, \boldsymbol{W}_e^T$ collaboratively learned in this MAR regression framework.

3.4. Optimization

The optimization problem Eq. (5) is solved by iteratively optimizing specific parameters at each step while holding the others fixed. The details are shown below:

Step 1. Fix $\boldsymbol{\theta}$ and optimize $\boldsymbol{W}$: the gradient of Eq. (5) with respect to $\boldsymbol{W}$ can be represented by $Gradient_w = \sum_{i=1}^{n}\frac{\partial||\boldsymbol{W}^T\widehat{\mathbf{X}}_i-\boldsymbol{Y}_i\theta||_{1,2}}{\partial\boldsymbol{W}}+\xi_s\boldsymbol{W}$, then the updated parameter at each time step is given by $\boldsymbol{W}(t)=\boldsymbol{W}(t-1)-\tau Gradient_w$. Here, τ is the iteration step size.

Step 2. Fix $\boldsymbol{W}$ and optimize θ: θ is solved by the standard gradient descent method. Specifically, the gradient of Eq. (5) with respect to θ is given by $Gradient_\theta=\sum_{i=1}^{n}\frac{\partial||\boldsymbol{W}^T\widehat{\mathbf{X}}_i-\boldsymbol{Y}_i\theta||_{1,2}}{\partial\theta}$, then the updated $\theta(t)=\theta(t-1)-\mu Gradient_\theta$ is projected into the constrained space. Here μ is the iteration step size.

Please note that $\boldsymbol{W}$ could be $\boldsymbol{W}_s$, $\boldsymbol{W}_p$, and $\boldsymbol{W}_e$, and $\boldsymbol{\theta}$ could be $\boldsymbol{\theta}_{sp}$, $\boldsymbol{\theta}_{spe}$, and $\boldsymbol{\theta}_{pe}$.

3.5. Activity fusion

In a practical video stream, it is difficult to decide the exact performance stage of a partially observed activity because of the ambiguous boundary among observations. Therefore, instead of using a single score function to identify it, we fuse the results from three score functions using a Gaussian function.

$$\arg\max_{label_n}\sum_{n=1}^{c}\sum_{k=s,p,e}G_k\boldsymbol{W}_k^T(label_n,:)\Phi(\mathbf{x}_i)\tag{6}$$

here,

$$G_k=\frac{1}{\sqrt{2\pi}}exp\left(-\frac{(\Phi(\mathbf{x}_i)-\mu_k)^2}{2\delta_k^2}\right)\tag{7}$$

where $\boldsymbol{W}_k^T(label_n,:)$ is the learned coefficients of category $label_n$, c is the number of activity classes, and G_k is the weight of the score produced by $\boldsymbol{W}_k^T(label_n,:)$. G_k is calculated by a Gaussian function with mean μ_k and standard deviation δ_k of observations at the performance stage $k(k\in(s,\ p,\ e))$.

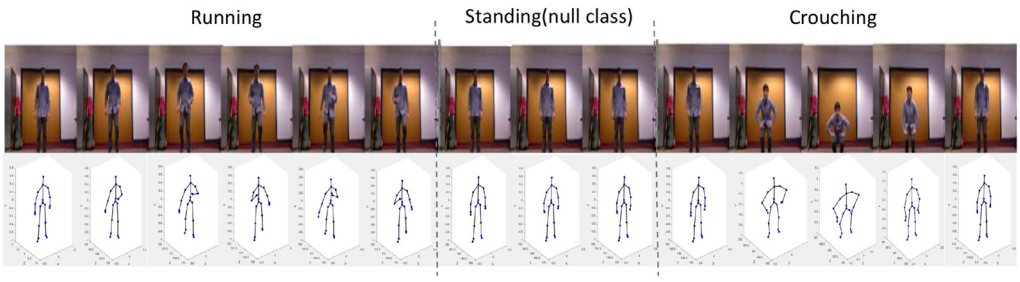
4. Experiment

4.1. Databases

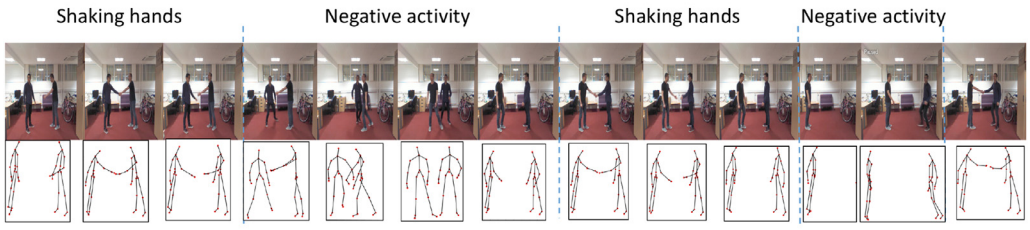
Multi-Modal Action Detection (MAD) database [19]. This is a sequential action database collected by Carnegie Mellon University in 2014. Fig. 4(a) reproduced from [19] shows some sample frames of the MAD database. The MAD database has 40 sequences performed by 20 subjects (2 sequences each subject). Each sequence contains 35 actions continuously performed by one subject. The time series between two actions are considered as the null class where the subject keeps standing in most cases. Three modalities: RGB videos, depth videos, and 3D coordinates of 20 skeleton joints are recorded using the Microsoft Kinect sensor.

Online Human Interaction (OHI) database. This database contains total 10 human interactions: *shaking hands, high waving, kicking, punching, pushing, hugging, high-fiving, approaching, departing and exchanging objects.* Each interaction is performed by 23 pairs of subjects for $2\sim4$ times. We record RGB images (640×480), depth images (640×480), and 3D coordinates of 20 skeleton joints for each pair of subjects using Kinect version 1. The registered depth images are further provided for the facility of fusing the information from both channels at the pixel level.

There are two parts in this database. In Part I, interactions are divided into isolated sequences according to interaction categories. This part is mainly designed for the offline activity recognition. For the evaluation of online activity recognition methods, Part II is collected where each video sequence contains 10 human interactions continuously performed by one pair of subjects. The ground truth labels along with the starting and ending points of activities are provided. During the interval of two activities, the subjects are free to perform any actions instead of standing still, which makes



(a) MAD database.



(b) OHI database.

Fig. 4. (a) An illustration of RGB and skeleton frames of the MAD database [19]. The time sequence between two actions is considered as *null class*. (b) An illustration of RGB and skeleton frames of the OHI database. The time sequences between two activities when subjects are performing casual behaviors are seen as negative activities.

this database closer to practical scenarios as well as challenging. Fig. 4(b) shows an example sequence of our database.

4.2. Experimental settings

The five-fold-cross-validation [19] is used for the experimental evaluation. Action sequences of four folds (4 subjects per fold) were used for training and the sequences of the remaining fold were used for evaluation. The experimental evaluation consists of two aspects: the performance over partial activity observations and comparative results in online activity recognition.

To systematically evaluate the MAR method in terms of the robustness of discriminating partial activity observations, a sliding window strategy is used to generate the partial activity fragments from complete activities in both databases. At training time, these activity fragments are further grouped into three performance stages (*start, peak* and *end*) based on their locations in the complete activity sequence. For each activity sequence, two manually annotated boundaries are provided to separate the three stages. If the end frame of an activity fragment is smaller than the first annotated boundary, it belongs to the *start* period; If it is within two boundaries, it belongs to the *peak*; otherwise, it belongs to the *end* period. It should be noted that although the annotation of the boundaries largely depends on human experience, the performance of MAR method will not be affected to a large extent since it employs the developed adaptive label strategy to automatically learn the relationship between adjacent performance stages. The criterion in this experiment is the identification accuracies on the partial activity observations from different performance stage.

The second experiment is for the evaluation of the performance in the online activity recognition scenario. The input is video sequences which contain multiple activities performed continuously. The online operation is implemented by using a sliding window with a length of 20 to sequentially select the activity segments over the continuous video streams. Two criteria *Precision percentage (Prec)* and *Recall percentage (Rec)* initially proposed in [19] are used for the evaluation: $Prec = \frac{TN}{DN}$ and $Rec = \frac{TN}{GTN}$. Where, *TN* represents the number of correctly detected activities who have 50% overlap with the ground truth activity; *DN* is the number of detected activities; *GTN* stands for the number of all ground truth activity classes. *Prec* and *Rec* indicate the accuracy of precisely detecting activity from videos and accurately classifying the detected activities, respectively.

4.3. Evaluation of MAR performance

This experiment intends to evaluate the performance of existing methods in terms of discriminating partial activity observations. Complete activities in both MAD and OHI databases are divided into three performance stages, from which overlapped activity segments are selected.

Table 1 shows the performance of the proposed methods on two databases. It can be found that, for the MR method, the recognition accuracies on the *start* and the *end* segments are much lower than the *peak* segments. This investigation verifies a fact that segments from the *peak* period are easier to be identified because their information is more discriminative than both *start* and *end* stages. The necessity of improving the performance of partial activity recognition, especially for segments from the *start* and *end* stages, motivates us to introduce the adaptive label strategy between adjacent stages in MAR to model the inherent evolution of activities. The comparison between MR and MAR is given to demonstrate the enhanced recognition ability of MAR on similar segments. The third row in Table 1 reports that the MAR method significantly outperforms the MR method by improving the average accuracies from 64.66% to 76.50% and from 47.46% to 63.76% on MAD database and OHI database, respectively. Specifically, for two databases, the recognition accuracies of MR on three stages are improved over 10% by MAR. More obviously, MAR achieves 60% on the *end* stage on the OHI database, which is approximately 20% higher than that of MR. Furthermore, the MAR method is also compared with the state-of-the-art RNN-based Beyond Joints+RNN [54]. Regarding the training of Beyond Joints+RNN, the publicly

Table 1
Recognition Accuracy (%) of partial activities from the MAD database and the OHI database.

	MAD database				OHI database			
	start	*peak*	*end*	Average	*start*	*peak*	*end*	Average
Beyond Joints+RNN [54]	69.40	75.89	64.11	**69.80**	49.61	69.90	56.40	**58.64**
MR	67.87	70.75	55.35	**64.66**	42.51	61.40	38.46	**47.46**
MAR	80.28	83.33	65.88	**76.50**	54.74	76.53	60.00	**63.76**

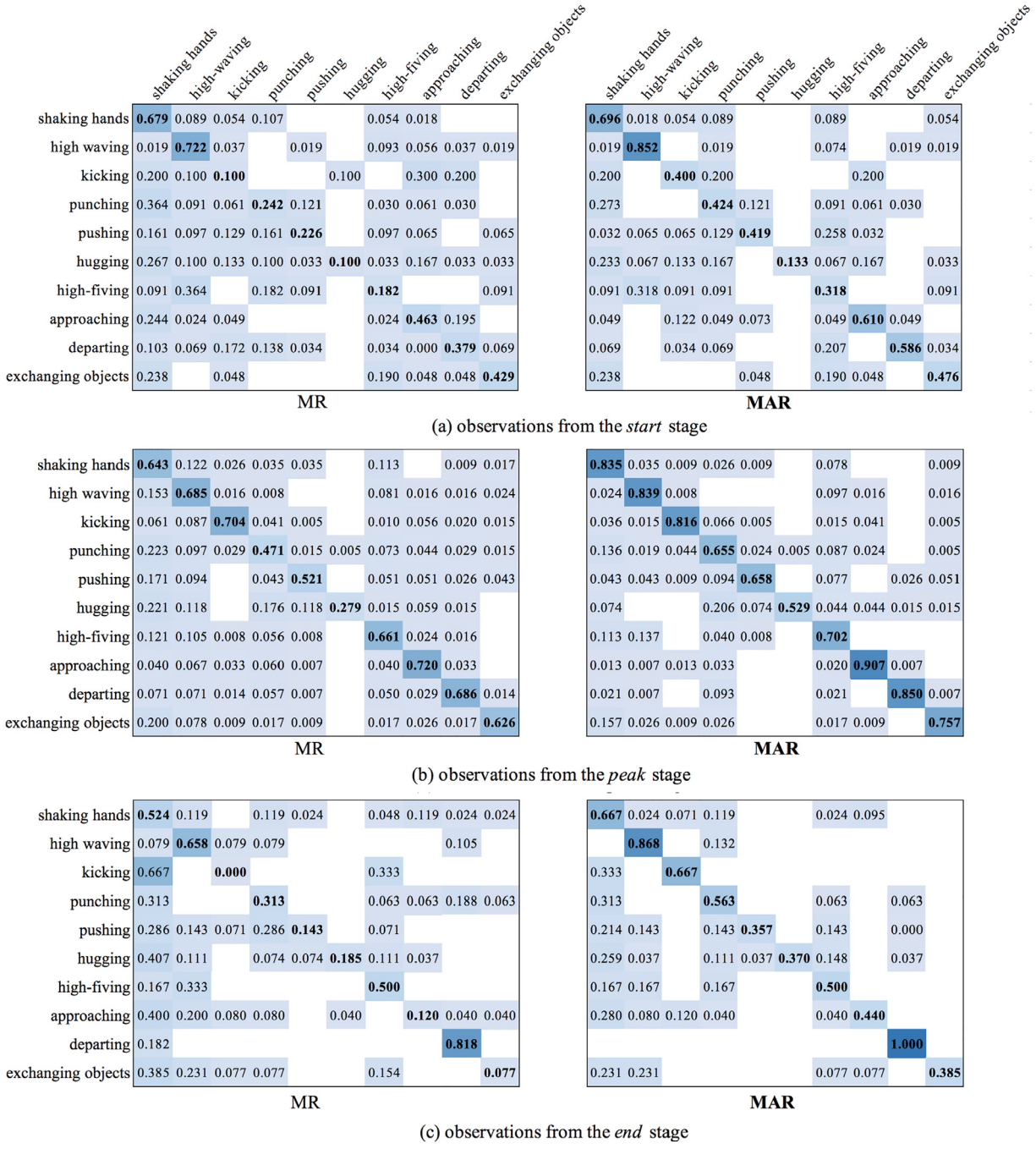


Fig. 5. Confusion matrixes of MR and MAR for observations from the *start*, *peak*, *end* performance stage on the OHI database.

available implementation[1] with slight modification towards skeleton joints and maximum training epoch is adopted to guarantee its performance. Since the number of skeleton joints in MAD dataset is different from the datasets used in [54], the model is directly trained from scratch. The proposed MAD method achieves better performance over Beyond Joints+RNN using the limited information provided by the partial activity observations. This achievement owes much to the collaboratively learned score functions with a

[1] https://github.com/hongsong-wang/Beyond-Joints .

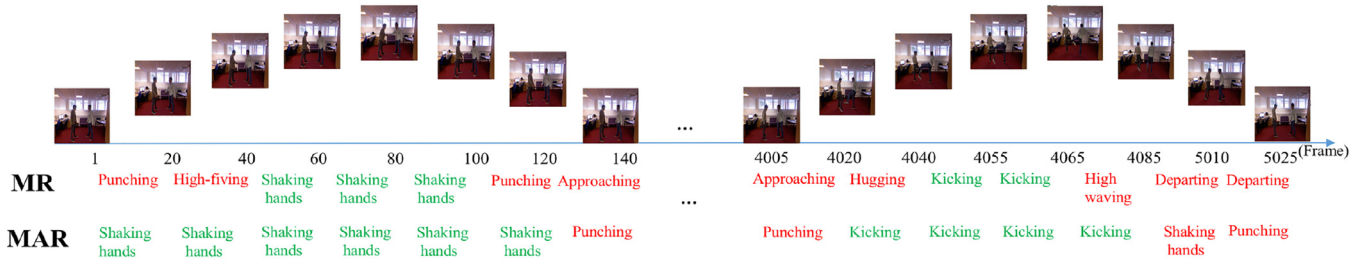


Fig. 6. The comparison between **MR** and **MAR** on a test sequence from OHI dataset. The results in green indicate the correct identified observations while the results in red indicates the wrong identified ones. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

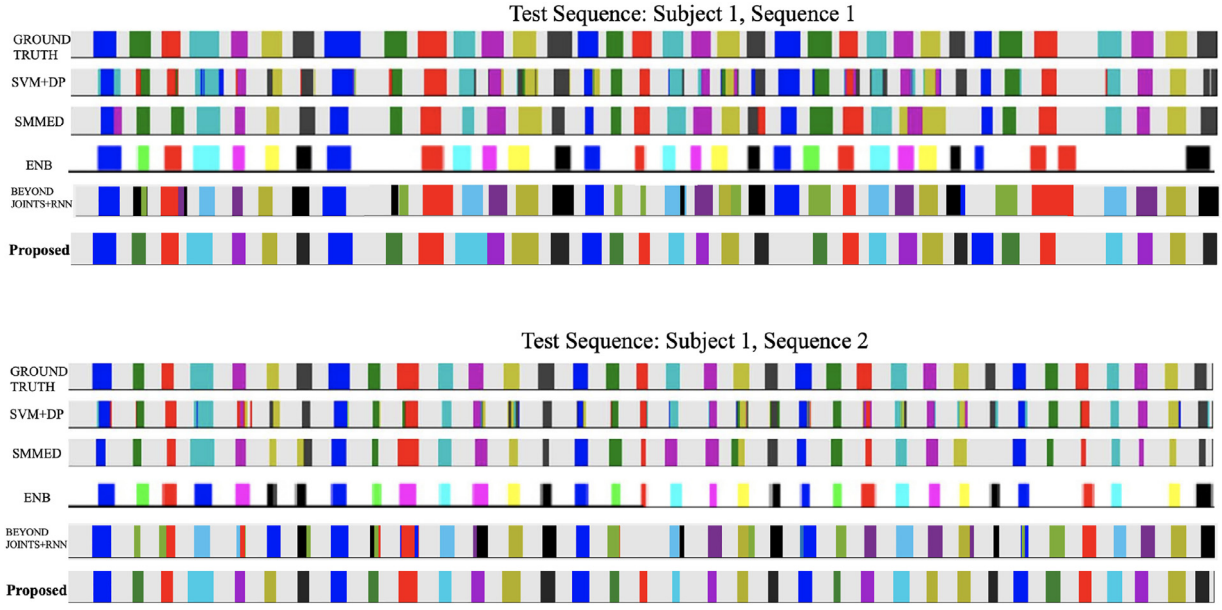


Fig. 7. The comparison of **MAR** with SVM+DP [59], SMMED [19], ENB [38] and Beyond Joints+RNN [54] on two test sequences from the MAD database.

focus on specific performance stages, which strengthens the recognition power of functions $\boldsymbol{W}_s^T$ and $\boldsymbol{W}_e^T$.

Fig. 5 shows the confusion matrixes of the MR and MAR method on the OHI database. In the MR method, the recognition accuracies of observations from the *start* and *end* stage in most activity classes are less than 50%, and even for the observations from the *peak* stage which posses relatively rich characteristics of each activity category. For example, activities such as *punching*, and *hugging* only have 47.1% and 27.9% accuracy mainly due to the confusion with *shaking hands*. This poor performance, especially for the beginning and ending parts, is mainly caused by the limited information from minor temporal activity sequences. The MAR method significantly boosts the overall recognition performance by learning a specific score function for each performance stage as well as considering the inherent evolution of activities via the adaptive label strategy. For example, MAR can correctly classify over 80% of the observations from the *peak* stage, and it dramatically improves the accuracy of *kicking* by 66.7%.

Fig. 6 reports the identification performance of the proposed MR and MAR method in the online activity recognition scenario. Two activity sequences from a random video stream in the OHI dataset are selected for analysis. A sliding window with a length of 20 is used to sequentially produce activity segments over the time. Both methods can correctly identify the activities in their peak period. However, accuracies of the MR method on the fragments at the beginning or ending of the activities are quite low due to the confusion with other activities, such as *shaking hands* is wrongly recognized as *punching* or *high-waving*. From the bottom line of Fig. 6, it can be seen that a significant improvement in discriminating the similar activity is achieved by the MAR framework.

Table 2
Online Performance: Average Detection and Recognition results on the MAD database(%).

Methods	*Prec*	*Rec*
SVM+DP [59]	28.6	51.4
SMMED [19]	59.2	57.4
ENB [38]	76.1	73.6
Method in [60]	72.1	79.7
Beyond Joints+RNN [54]	74.2	73.4
MAR	**81.3**	**82.3**

4.4. Comparative results in online activity recognition

This section aims to evaluate the performance of existing methods in the online activity recognition scenario. Following the setting in [19], the *Precision percentage (Prec)* and *Recall percentage (Rec)* are employed for the experimental evaluation in Table 2. As with the previous setting, the publicly available implementation[1] is used for the training. The parameters of maximum training epoch, skeleton joints, and window size are adjusted to achieve better performance. The model of Beyond Joints + RNN is directly trained from scratch using the temporal video sequence. The relatively low accuracies of the Beyond Joints + RNN method might due to its overfitting in the small dataset. On the other hand, the MAR method achieves the best performance of over 80% accuracy

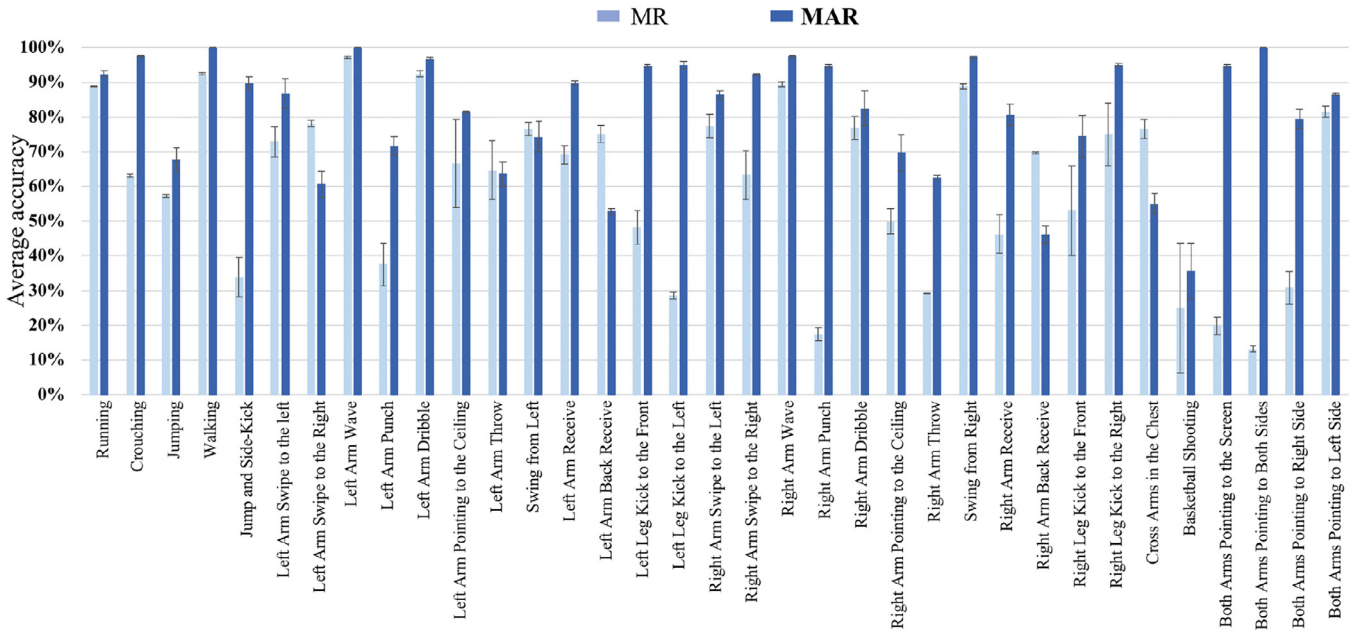


Fig. 8. Average recognition accuracy on each activity category. Error bars indicate standard deviations.

in both *Prec* and *Rec* among all methods, which demonstrates its capacity of precisely detecting activities from videos and accurately classifying the detected activities.

Fig. 7 shows the online performance on two commonly used test sequences. For each test sequence, the results of SVM+DP [59], SMMED [19], ENB [38], Beyond Joints + RNN [54] and MAR are listed to compare their performance. It can be seen that although SVM+DP and Beyond Joints + RNN can detect the occurrence of each activity in two test sequences, they have frequently fragmented labels around ground truth labels, indicating the low recognition accuracy on the detected activities. Though SMMED and ENB improve the recognition performance, they suffer from the high missing detection rate where the occurrence of several activities is not detected. On the other hand, MAR is capable of correctly detecting and classifying most of the activities. Some of the detected actions are misclassified due to their high similarities during a short interval. For example, the action *Basketball Shooting* has a short term period of the static pose which is quite similar to the pose in *Running*. Since the designed feature descriptor simultaneously considers dynamic trajectory and static poses, the missing of the dynamic movement in that short period might lead to misclassification.

Fig. 8 reports the average recognition accuracy of each activity category. The figure shows that MAR outperforms MR on most of the activities. Especially, MAR achieves significant improvements on *Both Arms Pointing to Right Side* and *Right Arm Punch*, by increasing accuracies from 13.13% to 100% and from 17.5% to 94.72%, respectively. Besides, compared to MR, MAR has smaller standard deviations of results in all activity categories. This demonstrates that the adaptability to random activity segments is strengthened by the collaboratively learned score functions in MAR.

5. Conclusion

This paper proposes a multi-stage adaptive regression framework to deal with the partial activity observation problem in online activity recognition. The MAR framework delicately assembles overlapped activity observations and also considers the relation between adjacent performance stages to improve its robustness to arbitrary activity segments. By formulating the online activity recognition task as a multi-stage adaptive regression problem, multiple score functions are collaboratively learned to effectively discriminate similar partial activity fragments.

The improvements of MAR over MR are firstly validated by over 10% increases of the accuracies in each performance stage on both MAD and OHI databases, owing to reducing the confusion among activities. Furthermore, the MAR method achieves an accuracy of 82.3% on the MAD database with a significant improvement of 8.7% over the state of the art. In addition, the online human interaction database is introduced to be served for the evaluation of online human interaction recognition methods. The database is realistic and challenging because interactions are performed continuously and the temporal sequences between two interactions include complex activities rather than the still status in existing databases.

While the proposed method has been successful in improving online human activity recognition, there remains an assumption that activities in the training dataset can be segmented into three different stages. Currently, a manual labeling procedure is employed to conduct the segmentation, which limits its applications on large datasets such as PKU-MMD [55]. Thus, one of our future work will be enhancing the algorithm with an automatic labeling procedure.

Although online activity recognition is essential in practical applications, it is still a problem far from being solved. This paper focuses on exploring the feasibility of solving partial activities via multi-stage adaptive regression rather than the construction of features. Thus, a promising future direction to further improve the performance will be exploring an effective integration of deep learning features into the proposed framework.

Acknowledgment

This work was supported in part by the EU Seventh Framework Programme (no. 611391, Development of Robot-Enhanced therapy for children with AutisM spectrum disorders (DREAM)).

References

[1] J. Aggarwal, L. Xia, Human activity recognition from 3d data: a review, Pattern Recognit. Lett. 48 (2014) 70–80.
[2] A. Marcos-Ramiro, D. Pizarro, M. Marron-Romera, D. Gatica-Perez, Let your body speak: communicative cue extraction on natural interaction using rgbd data, IEEE Trans. Multimedia 17 (10) (2015) 1721–1732.
[3] B. Liu, H. Cai, Z. Ju, H. Liu, Rgb-d sensing based human action and interaction analysis: a survey, Pattern Recognit. 94 (2019) 1–12.
[4] O.D. Lara, M.A. Labrador, et al., A survey on human activity recognition using wearable sensors., IEEE Commun. Surv. Tuts. 15 (3) (2013) 1192–1209.
[5] S. Althloothi, M.H. Mahoor, X. Zhang, R.M. Voyles, Human activity recognition using multi-features and multiple kernel learning, Pattern Recognit. 47 (5) (2014) 1800–1812.
[6] O. Oreifej, Z. Liu, Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2013, pp. 716–723.
[7] M. Zanfir, M. Leordeanu, C. Sminchisescu, The moving pose: an efficient 3d kinematics descriptor for low-latency action recognition and detection, in: Proc. IEEE Int. Conf. Comput. Vis., 2013, pp. 2752–2759.
[8] Y. Guo, Y. Li, Z. Shao, Dsrf: a flexible trajectory descriptor for articulated human action recognition, Pattern Recognit. (2017).
[9] A. Jalal, Y.-H. Kim, Y.-J. Kim, S. Kamal, D. Kim, Robust human activity recognition from depth video using spatiotemporal multi-fused features, Pattern Recognit. 61 (2017) 295–308.
[10] M. Liu, H. Liu, C. Chen, Robust 3d action recognition through sampling local appearances and global distributions, IEEE Trans. Multimedia (2017).
[11] S. Zhang, Y. Yang, J. Xiao, X. Liu, Y. Yang, D. Xie, Y. Zhuang, Fusing geometric features for skeleton-based action recognition using multilayer lstm networks, IEEE Trans. Multimedia (2018).
[12] Y. Shi, Y. Tian, Y. Wang, T. Huang, Sequential deep trajectory descriptor for action recognition with three-stream cnn, IEEE Trans. Multimedia 19 (7) (2017) 1510–1520.
[13] C. Wang, Z. Liu, S.-C. Chan, Superpixel-based hand gesture recognition with kinect depth camera, IEEE Trans. Multimedia 17 (1) (2015) 29–39.
[14] R. Qiao, L. Liu, C. Shen, A. van den Hengel, Learning discriminative trajectorylet detector sets for accurate skeleton-based action recognition, Pattern Recognit. 66 (2017) 202–212.
[15] M. Ryoo, Human activity prediction: Early recognition of ongoing activities from streaming videos, in: Proc. IEEE Int. Conf. Comput. Vis., 2011, pp. 1036–1043.
[16] T. Lan, T.-C. Chen, S. Savarese, A hierarchical representation for future action prediction, in: Proc. Eur. Conf. Comput. Vis., 2014, pp. 689–704.
[17] Y. Kong, D. Kit, Y. Fu, A discriminative model with multiple temporal scales for action prediction, in: Proc. Eur. Conf. Comput. Vis., 2014, pp. 596–611.
[18] M. Hoai, F. De la Torre, Max-margin early event detectors, Int. J. Comput. Vis. 107 (2) (2014) 191–202.
[19] D. Huang, S. Yao, Y. Wang, F. De La Torre, Sequential max-margin event detectors, in: Proc. Eur. Conf. Comput. Vis., 2014, pp. 410–424.
[20] X. Cai, W. Zhou, L. Wu, J. Luo, H. Li, Effective active skeleton representation for low latency human action recognition, IEEE Trans. Multimedia 18 (2) (2016) 141–154.
[21] X. Wang, L. Gao, P. Wang, X. Sun, X. Liu, Two-stream 3-d convnet fusion for action recognition in videos with arbitrary size and length, IEEE Trans. Multimedia 20 (3) (2018) 634–644.
[22] F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal, R. Bajcsy, Sequence of the most informative joints (smij): a new representation for human skeletal action recognition, J. Vis. Commun. Image Represent. 25 (1) (2014) 24–38.
[23] R. Vemulapalli, F. Arrate, R. Chellappa, Human action recognition by representing 3d skeletons as points in a lie group, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2014, pp. 588–595.
[24] C. Ellis, S.Z. Masood, M.F. Tappen, J.J. Laviola Jr, R. Sukthankar, Exploring the trade-off between accuracy and observational latency in action recognition, Int. J. Comput. Vis. 101 (3) (2013) 420–436.
[25] Y. Guo, Y. Li, Z. Shao, Dsrf: a flexible trajectory descriptor for articulated human action recognition, Pattern Recognit. 76 (2018) 137–148.
[26] I. Lillo, J.C. Niebles, A. Soto, Sparse composition of body poses and atomic actions for human activity recognition in rgb-d videos, Image Vis. Comput. 59 (2017) 63–75.
[27] X. Ji, J. Cheng, W. Feng, D. Tao, Skeleton embedded motion body partition for human action recognition using depth sequences, Signal Process. 143 (2018) 56–68.
[28] Y. Yang, C. Deng, S. Gao, W. Liu, D. Tao, X. Gao, Discriminative multi-instance multitask learning for 3d action recognition, IEEE Trans. Multimedia 19 (3) (2017) 519–529.
[29] S. Yan, Y. Xiong, D. Lin, Spatial temporal graph convolutional networks for skeleton-based action recognition, arXiv:1801.07455 (2018).
[30] J. Liu, N. Akhtar, A. Mian, Skepxels: spatio-temporal image representation of human skeleton joints for action recognition, arXiv:1711.05941 (2017).
[31] J. Liu, G. Wang, L.-Y. Duan, K. Abdiyeva, A.C. Kot, Skeleton-based human action recognition with global context-aware attention lstm networks, IEEE Trans. Image Process. 27 (4) (2018) 1586–1599.
[32] Y. Hou, Z. Li, P. Wang, W. Li, Skeleton optical spectra based action recognition using convolutional neural networks, IEEE Trans. Circuits Syst. Video Technol. 28 (3) (2018) 807–811.
[33] Q. Ke, M. Bennamoun, S. An, F. Sohel, F. Boussaid, A new representation of skeleton sequences for 3d action recognition, in: Proc. Conf. Comput. Vis. Pattern Recognit., 2017, pp. 4570–4579.
[34] Y. Du, W. Wang, L. Wang, Hierarchical recurrent neural network for skeleton based action recognition, in: Proc. Conf. Comput. Vis. Pattern Recognit., 2015, pp. 1110–1118.
[35] V. Veeriah, N. Zhuang, G.-J. Qi, Differential recurrent neural networks for action recognition, in: Proc. IEEE Int. Conf. Comput. Vis., 2015, pp. 4041–4049.
[36] A. Shahroudy, J. Liu, T.-T. Ng, G. Wang, Ntu rgb+ d: a large scale dataset for 3d human activity analysis, in: Proc. Conf. Comput. Vis. Pattern Recognit., 2016, pp. 1010–1019.
[37] Y. Du, Y. Fu, L. Wang, Representation learning of temporal dynamics for skeleton-based action recognition, IEEE Trans. Image Process. 25 (7) (2016) 3010–3022.
[38] H.J. Escalante, E.F. Morales, L.E. Sucar, A naive Bayes baseline for early gesture recognition, Pattern Recognit. Lett. 73 (2016) 91–99.
[39] J.-F. Hu, W.-S. Zheng, L. Ma, G. Wang, J. Lai, Real-time rgb-d activity prediction by soft regression, in: Proc. Eur. Conf. Comput. Vis., 2016, pp. 280–296.
[40] R. De Geest, E. Gavves, A. Ghodrati, Z. Li, C. Snoek, T. Tuytelaars, Online action detection, in: Proc. Eur. Conf. Comput. Vis., 2016, pp. 269–284.
[41] J.-L. Reyes-Ortiz, L. Oneto, A. Samà, X. Parra, D. Anguita, Transition-aware human activity recognition using smartphones, Neurocomputing 171 (2016) 754–767.
[42] K. Kulkarni, G. Evangelidis, J. Cech, R. Horaud, Continuous action recognition based on sequence alignment, Int. J. Comput. Vis. 112 (1) (2015) 90–114.
[43] C. Wu, J. Zhang, S. Savarese, A. Saxena, Watch-n-patch: Unsupervised understanding of actions and relations, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2015, pp. 4362–4370.
[44] S. Nowozin, J. Shotton, Action points: a representation for low-latency online human action recognition, Microsoft Research Cambridge, Tech. Rep. MSR-TR-2012-68 (2012).
[45] S. Fothergill, H. Mentis, P. Kohli, S. Nowozin, Instructing people for training gestural interactive systems, in: Proc. SIGCHI Conf. Human Factors in Comput. Syst., 2012, pp. 1737–1746.
[46] V. Bloom, V. Argyriou, D. Makris, Dynamic feature selection for online action recognition, in: Int. Workshop Human Behavior Understanding, 2013, pp. 64–76.
[47] A. Sharaf, M. Torki, M.E. Hussein, M. El-Saban, Real-time multi-scale action detection from 3d skeleton data, in: IEEE Winter Conf. Appl. Comput. Vision, 2015, pp. 998–1005.
[48] V. Bloom, V. Argyriou, D. Makris, Linear latent low dimensional space for online early action recognition and prediction, Pattern Recognit. 72 (2017) 532–547.
[49] P. Molchanov, X. Yang, S. Gupta, K. Kim, S. Tyree, J. Kautz, Online detection and classification of dynamic hand gestures with recurrent 3d convolutional neural network, in: Proc. Conf. Comput. Vis. Pattern Recognit., 2016, pp. 4207–4215.
[50] S. Song, C. Lan, J. Xing, W. Zeng, J. Liu, Spatio-temporal attention based lstm networks for 3d action recognition and detection, IEEE Trans. Image Process. pp (99) (2018). 1–1
[51] Y. Li, C. Lan, J. Xing, W. Zeng, C. Yuan, J. Liu, Online human action detection using joint classification-regression recurrent neural networks, in: Proc. Eur. Conf. Comput. Vis., 2016, pp. 203–220.
[52] Z. Shou, D. Wang, S. Chang, Action temporal localization in untrimmed videos via multi-stage CNNS, in: Proc. Conf. Comput. Vis. Pattern Recognit., 3, 2016.
[53] X. Chai, Z. Liu, F. Yin, Z. Liu, X. Chen, Two streams recurrent neural networks for large-scale continuous gesture recognition, in: Int. Conf. Pattern Recognit. Workshops, 2016, pp. 31–36.
[54] H. Wang, L. Wang, Beyond joints: learning representations from primitive geometries for skeleton-based action recognition and detection, IEEE Trans. Image Process. 27 (9) (2018) 4382–4394.
[55] C. Liu, Y. Hu, Y. Li, S. Song, J. Liu, Pku-mmd: a large scale benchmark for continuous multi-modal human action understanding, arXiv:1703.07475 (2017).
[56] A. Graves, S. Fernández, F. Gomez, J. Schmidhuber, Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks, in: Proc. 23rd Int. Conf. Mach. Learn., ACM, 2006, pp. 369–376.
[57] J. Shotton, T. Sharp, A. Kipman, A. Fitzgibbon, M. Finocchio, A. Blake, M. Cook, R. Moore, Real-time human pose recognition in parts from single depth images, Commun. ACM 56 (1) (2013) 116–124.
[58] B. Liu, H. Yu, X. Zhou, D. Tang, H. Liu, Combining 3d joints moving trend and geometry property for human action recognition, in: IEEE Int. Conf. Syst. Man, Cyber., 2016, pp. 000332–000337.
[59] M. Hoai, Z.-Z. Lan, F. De la Torre, Joint segmentation and classification of human actions in video, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2011, pp. 3265–3272.
[60] M. Devanne, S. Berretti, P. Pala, H. Wannous, M. Daoudi, A. Del Bimbo, Motion segment decomposition of rgb-d sequences for human behavior understanding, Pattern Recognit. 61 (2017) 222–233.

Bangli Liu received the B.Eng. degree at the China Jiliang University in 2012 and the M.Sc. degree at East China University of Science and Technology in 2015, and the Ph.D. degree in computer science at the University of Portsmouth, UK, in 2018. She is currently a Research Associate in Loughborough University. Her research interests include computer vision, pattern recognition, machine learning, image super-resolution, human motion analysis.

Haibin Cai received the M.Sc. degree at the Zhejiang University of Technology in 2015 and the Ph.D. degree at the University of Portsmouth in 2018. He is currently a Research Associate in Loughborough University. His research interests include gaze estimation, motion recognition, object detection and tracking.

Zhaojie Ju (M'08, SM'16) received the B.S. in automatic control and the M.S. in intelligent robotics both from Huazhong University of Science and Technology, China, in 2005 and 2007 respectively, and the Ph.D. degree in intelligent robotics at the University of Portsmouth, UK, in 2010. He is currently a Senior Lecturer in the School of Computing, University of Portsmouth. He previously held research appointments at University College London and University of Portsmouth, UK. His research interests include machine intelligence, pattern recognition, and their applications on human motion analysis, human-robot interaction and collaboration, and robot skill learning. He has authored or co-authored over 100 publications in journals, book chapters, and conference proceedings and received four best paper awards. Dr. Ju is an Associate Editor of the IEEE Transactions on Cybernetics.

Honghai Liu (M'02, SM'06) received the Ph.D. in intelligent robotics from King's College London, London, U.K., in 2003. He is currently Chair Professor of intelligent systems and Robotics. His research interests include wearable sensing, biomechatronics, intelligent video analytics, intelligent robotics, and their practical applications with an emphasis on approaches that could make contribution to the intelligent connection of perception to action using contextual information. He is Associate Editor of IEEE Transactions on Human Machine Systems, IEEE Transactions on Industrial Informatics and IEEE Transactions on Industrial Electronics.