FT-HID: a large-scale RGB-D dataset for first- and third-person human interaction analysis

Zihui Guo1 ° · Yonghong Hou1 · Pichao Wang4 · Zhimin Gao2 · Mingliang Xu2 · Wanqing Li3

Received: 28 February 2022 / Accepted: 6 September 2022 / Published online: 7 October 2022
© The Author(s), under exclusive licence to Springer-Verlag London Ltd., part of Springer Nature 2022

Abstract
Analysis of human interaction is one important research topic of human motion analysis. It has been studied either using first-person vision (FPV) or third-person vision (TPV). However, the joint learning of both types of vision has so far attracted little attention. One of the reasons is the lack of suitable datasets that cover both FPV and TPV. In addition, existing benchmark datasets of either FPV or TPV have several limitations, including the limited number of samples, participant subjects, interaction categories, and modalities. In this work, we contribute a large-scale human interaction dataset, namely FT-HID dataset. FT-HID contains pair-aligned samples of first-person and third-person visions. The dataset was collected from 109 distinct subjects and has more than 90K samples for three modalities. The dataset has been validated by using several existing action recognition methods. In addition, we introduce a novel multi-view interaction mechanism for skeleton sequences, and a joint learning multi-stream framework for first-person and third-person visions. Both methods yield promising results on the FT-HID dataset. It is expected that the introduction of this vision-aligned large-scale dataset will promote the development of both FPV and TPV, and their joint learning techniques for human action analysis.

Keywords Human interaction · First-person vision · Third-person vision · Large-scale dataset

1 Introduction

Human action recognition is a fundamental yet challenging task in the field of computer vision. It has a wide range of practical applications including security surveillance, medical care, description generation, video retrieval, and human–robot interaction. According to the different emphasis of the analysis, the problem can be generalized into three pairs of learning problems, namely third-person vision (TPV) learning and first-person vision (FPV) learning, single-view learning and multi-view learning, as well as single-person action learning and multi-person interaction learning. Among them, TPV is the most common data paradigm in current action recognition research [21, 30, 48, 92]. When capturing TPV data, the camera can be seen as an observer. The subjects perform the action spontaneously without the intention to interact with the camera. Usually, the TPV camera is fixed at a distance of the subject to get stable and complete action and background environment information. There are many established RGB-D TPV datasets, such as NTU RGB-D dataset...
[57], SYSU dataset [22], UTD dataset [9], and M2I dataset [81]. However, these datasets still have limitations in some important aspects, such as the number of subjects, background environment, and the number of action categories, which are crucial for building a realistic yet challenging action recognition dataset. For example, SYSU 3D Human–Object interaction dataset [22] only contains 12 distinct actions performed by four subjects, and the large-scale NTU dataset [57] lacks data taken from the top view. The small number of camera views and less characteristic of the subjects may significantly limit the intra-class variation of actions, while the small set of distinguishable classes can be easily recognized and hinder real-world applications.

Different from TPV learning, the camera is a part of the performer in the collection of FPV data, which is obtained by the camera mounted on a wearable device of the performer. FPV provides the inherent human-centric perspective of human action and has unique characteristics, including the egocentric motion made by the movement of the head of the performer, the detailed information caused by the short distance from the camera to action target, and the narrow field-of-view of the FPV camera. Compared with TPV RGB-D action recognition, limited research has been conducted in the FPV domain. The main reason is the scarcity of the public available RGB-D egocentric dataset. To the best of our knowledge, only five existing FPV datasets [20, 50, 65, 79, 89] contain both RGB and depth modality. These datasets also have the same limitations as TPV datasets in the number of subjects, action category, and background environment. Moreover, most of the mentioned datasets focus on hand action such as gesture and hand–object interaction. However, the actions completed by other body parts as well as the multi-person interaction is also important in daily life.

TPV and FPV data have their unique strengths and weaknesses, respectively. Generally, TPV data are often stationary or moving smoothly, and stays off the subject, which can get a wide range of perspectives and capture more background information, but it lacks the details of the action. FPV data is close to the real scene that the wearer sees, and can record some details that cannot be seen from TPV. However, it gets a limited field of vision and loses some holistic information. Therefore, it is necessary to combine the two types of visions for better human action understanding and analysis. However, most existing datasets only have one kind of vision. For example, the datasets MSR Action 3D [40], UTDMHAD [9] and NTU RGB-D [57] only contain TPV data, while the datasets ADL [54], GTEA [13] and multimodal egocentric activity [63] are pure FPV datasets. Although the dataset CMU-MMAC [13] contains both visions, it lacks the depth modality.

Single-view learning is the conventional approach, and it is widely explored in existing research. As all the data are recorded in a static camera view, the core problem of single-view learning is that it does not consider the difference in appearance and motion characteristics from different camera views which may limit the diversity of action expressions and greatly affects the generalization of the recognition algorithm. Besides, some actions are difficult to recognize from certain viewpoints but easy to recognize from others. Thus multi-view learning, which aims to map features obtained from multiple views into a common feature space to handle the variations in visual appearance, is in aid of the action recognition field. A few existing datasets [70, 71, 75] set multiple cameras from different views but most of them only contain horizontal views and omit the vertical view. For instance, the view in UESTC dataset [27] cover 360°, but they are all at the same height. However, the data from the top view can provide unique information of the action and is especially important for intelligent video surveillance.

Single-person learning includes recognizing the periodic actions performed by only one subject, like walking and waving, and the person–object interaction such as picking a ball and sending a card. Extensive previous research and datasets [5, 38, 78, 86] focused on single-person action recognition. However, human interactions such as pushing and handshaking are more typical human activities that occur in public places. Human interaction usually involves at least two individual motions from multiple persons, who are concurrently interrelated with each other. Recognition of complex interactions between multiple persons will be necessary for a number of applications, including automatic detection of violent activities in smart surveillance systems. The human interactions are coordinated in the sense that the movement of one person depends on the movement of the other and vice versa, for example, person A kicks person B, and then person B strikes back. In contrast to single-person motions, the study of interactions between multiple persons could help extend the understanding of human motion. Two of the most limitation of existing interaction datasets [66, 77, 85] are the shortage of interaction categories and the number of samples in each category.

From the above analysis, it can be seen that pairs of complementary TPV and FPV data can represent the action more comprehensively. Meanwhile, compared with single-view and single-person learning, multi-view and multi-person interaction learning are more challenging and meaningful. However, as shown in Table 1, the current datasets cannot meet the above three research demands simultaneously and also remain defects in each of the three learning tasks. Therefore, in this paper, a large-scale RGB-D dataset (FT-HID dataset) for first- and third-person
vision analysis with multiple camera views and diverse multi-person interactions is collected to fill the gap. The dataset contains more than 38K RGB samples, 38K depth samples, and about 20K skeleton sequences. Thirty classes of daily actions are designed specifically for multi-person interaction with a wearable device and three fixed cameras. FT-HID dataset has a comparable number of data, action classes, and scenes with other RGB-D action recognition datasets. It is more complex as the data are collected from 109 distinct subjects with large variations in gender, age, and physical condition. More importantly, to the best of our knowledge, it is the first large-scale RGB-D dataset that is collected from both TPV and FPV perspectives for action recognition.

Given the FT-HID dataset, we evaluate several published methods based on both handcrafted and deep learning features on TPV action recognition and FPV action recognition separately with suggested cross-subject and cross-view evaluation criteria. Furthermore, to explore the intrinsic relation between TPV and FPV, the performance of fusion across two visions is also investigated. It is believed that the FT-HID dataset can be used as a benchmark and help the community to move steps forward in the exploration of both FPV and TPV action recognition.

Moreover, two novel deep neural networks are proposed for action recognition. One aims to adequately learn the nonlinear structure of heterogeneous representations from different modalities and views and to exploit their complementary characteristics. Specifically, four 3D deep convolutional neural networks are utilized to learn the spatiotemporal information of four streams, i.e., TPV RGB videos, TPV depth sequences, FPV RGB videos, and FPV depth sequences, respectively. The other one focuses on exploring the joint representation of skeleton sequences from different camera views. It is realized by integrating channel-selection modules into a two-stream structure.

| Dataset                                | TPV | FPV | Total-action | Interaction | Subject | Camera | Video-sample | Year |
|----------------------------------------|-----|-----|--------------|-------------|---------|--------|--------------|------|
| WVCS dataset [50]                      | ×   | ✓   | 18           | 0           | 4       | 1      | 918          | 2014 |
| Robot-Centric Activity dataset [79]    | ×   | ✓   | 9            | 9           | 8       | 1      | 189          | 2015 |
| First-Person Hand Action dataset [20]  | ×   | ✓   | 45           | 0           | 6       | 1      | 1175         | 2018 |
| EgoGesture dataset [89]                | ✓   | ×   | 83           | 0           | 50      | 1      | 24,161       | 2018 |
| THU-READ dataset [65]                  | ✓   | ×   | 16           | 0           | 8       | 1      | 1920         | 2018 |
| MSR Daily Activity 3D dataset [69]     | ×   | ✓   | 10           | 0           | 10      | 1      | 320          | 2012 |
| UTKinect dataset [78]                  | ✓   | ×   | 20           | 0           | 10      | 1      | 200          | 2012 |
| G3D dataset [3, 5]                     | ✓   | ×   | 5            | 0           | 5       | 1      | 600          | 2012 |
| Falling Event dataset [86]             | ✓   | ×   | 5            | 0           | 5       | 1      | 150          | 2012 |
| SBU dataset [85]                       | ✓   | ×   | 8            | 8           | 7       | 1      | 300          | 2012 |
| Osaka dataset [49]                     | ✓   | ×   | 10           | 0           | 8       | 1      | 80           | 2012 |
| WorkoutSU-10 dataset [51]              | ✓   | ×   | 10           | 0           | 12      | 1      | 1200         | 2013 |
| Multi-view 3D Event dataset [75]       | ✓   | ×   | 8            | 0           | 8       | 3      | 3815         | 2013 |
| MAD dataset [24]                       | ✓   | ×   | 35           | 0           | 20      | 1      | 1400         | 2014 |
| HARL dataset [77]                      | ✓   | ×   | 10           | 3           | 21      | –      | 828          | 2014 |
| Office Activity dataset [71]           | ✓   | ×   | 20           | 10          | 10+     | 3      | 1180         | 2014 |
| ShakeFive dataset [66]                 | ✓   | ×   | 2            | 2           | 37      | 1      | 100          | 2014 |
| G3Di [4]                               | ✓   | ×   | 18           | 18          | 12      | 1      | –            | 2014 |
| N-UCLA dataset [70]                    | ✓   | ×   | 10           | 0           | 10      | 3      | 1475         | 2014 |
| UWA3D dataset [55]                     | ✓   | ×   | 30           | 0           | 10      | –      | 600+         | 2014 |
| TJIU dataset [42]                      | ✓   | ×   | 15           | 0           | 20      | 1      | 1200         | 2015 |
| UTD dataset [9]                        | ✓   | ×   | 27           | 0           | 8       | 1      | 861          | 2015 |
| MV-TIU dataset [44]                    | ✓   | ×   | 22           | 0           | 20      | 2      | 3520         | 2015 |
| NTU-60 dataset [57]                    | ✓   | ×   | 60           | 11           | 40      | 3      | 56,880       | 2016 |
| M²I dataset [43]                       | ✓   | ×   | 22           | 9            | 22      | 2      | 1784         | 2017 |
| UESTC RGB-D dataset [27]               | ✓   | ×   | 40           | 0           | 118     | 8      | 25,600       | 2018 |
| NTU-120 dataset [47]                   | ✓   | ×   | 120          | 26           | 106     | 3      | 114,480      | 2019 |
| FT-HID dataset                         | ✓   | ✓   | 30           | 30           | 109     | 4      | 38,364       | 2021 |
2 Related work

In this section, a brief overview of the current available RGB-D action recognition datasets is proposed. Then, some handcrafted and deep learning-based methods are discussed. For both the datasets and methods, TPV and FPV conditions are introduced separately.

2.1 Datasets

2.1.1 TPV dataset

In the field of human action recognition, most of the established datasets are captured in TPV. Among all the TPV datasets, the MSRDailyActivity3D dataset [69], UTKinect dataset [78], UTD dataset [9], NTU-60 [57], and NTU-120 [47] are commonly used benchmarks. MSRDailyActivity3D dataset [69] consists of 320 samples of 16 action classes, which are performed by ten actors in two background settings. UTKinect dataset [78] contains 200 samples collected from ten subjects by Kinect sensor. It collects ten classes of daily actions in total. In UTD dataset [9], eight subjects perform 27 actions in a fixed background environment and obtains 861 samples. These three benchmarks contain only single-person actions and are captured with only one camera at a time. In addition, the number of subjects participating in the performance and the overall sample size are also very limited. NTU provides a larger dataset [57] of 60 actions from 40 distinct subjects. Its new version [47] extends the action category into 120 and the total number of samples is 114,480. Although three cameras are used to capture the same action at the same time in NTU datasets [47, 57], only the horizontal views are included, and the vertical view is missing. In addition to the five benchmarks, all the other datasets also contain one or more limitations, including the small number of action categories, interactions, subjects, camera views, and sample size.

2.1.2 FPV dataset

The number of available FPV datasets is smaller by an order of magnitude than their TPV equivalents. One of the reasons is the difficulty of data collection. Most existing wearable cameras such as GoPro, Goole Glass, and Looxcie only capture RGB videos. Besides, the size of some RGB-D cameras including Kinect, Intel Real Sense, and ZED are relatively big and need real-time computer control to measure depth range which makes them unsuitable for egocentric data collection. There exist only five RGB-D datasets for FPV action recognition. WCVS dataset [50] is the earliest work that opened up the research in RGB-D FPV action recognition. Robot-Centric Activity dataset [79] is collected from a Kinect device mounted on top of an autonomous non-humanoid robot. First-Person Hand Action dataset [20] focuses on daily hand–object interaction and the RGB-D camera is mounted on the shoulder of each subject. EgoGesture dataset [89] is designed for egocentric hand gesture recognition which can be seen as a fine-grained action recognition task. Notably, the subjects are taught how to perform each gesture before data collection, which may decrease the inter-class variation and the reality of the dataset. THU-READ dataset [65] contains 1920 video clips with 40 hand actions. During the data collection, the camera is kept in the same direction as the subject’s eyesight. The most obvious limitation of the current FPV datasets is the small number of interaction classes and sample size.

2.2 Action recognition methods

2.2.1 TPV methods

Since the RGB videos can maintain color and texture information while the depth maps are invariant to illumination changes, many methods [19, 33, 58, 74, 88] have been proposed to fuse features extracted from both depth and RGB modalities. The handcrafted descriptors, including HOG and HOF, are widely used in early RGB-D action recognition approaches [28, 32, 34, 35, 45]. In addition, some work focus on designing and capturing suitable features to describe the trajectories [1, 17, 46, 64] and motions [18, 25, 26] of the action. Specifically, Asadi et al. [1] propose multi-modal dense trajectories to describe RGB-D videos and utilize a hybrid median filter as well as inpainting technique to enhance depth videos. Based on the dense trajectory features and human pose representations extracted from RGB and depth videos, Gao et al. [17] propose a collaborative sparse representation learning model for RGB-D action recognition. Imran et al. [26] train a four-channel deep convolutional neural network to fuse motion history images and depth motion maps extracted from RGB and depth videos, respectively. Besides, binary local representation is adopted in [83, 84] to preserve the pairwise structure with shape constraints through an orthogonal projection matrix. Except for RGB videos and depth maps, skeleton sequences are also involved in some
work [22, 23, 31, 56] to enhance the recognition performance. However, skeleton-based action recognition without RGB video and depth maps can also achieve satisfactory performance due to the advantages of skeleton sequences with detailed position and motion information, insensitivity to the environment, and low storage space.

Existing studies focus on using skeleton data directly to achieve skeleton-based action recognition mainly targets with RNN [2, 12, 68, 88], CNN [6, 52, 67], and GCN [16, 36, 39, 60, 61, 76]. RNN has a strong ability to model sequential data which is consistent with the intrinsic structure of the skeletal sequence and generally takes skeleton joint coordinates vector as the input of networks. Zhang et al. [88] enhance the RNN through a universal spatial modeling method, then a multi-stream LSTM with a score fusion technique is proposed to combine multiple geometric features. The CNN-based methods convert the skeleton data into pseudo-images or joint trajectory maps for simulating the RGB maps. For example, Cao et al. [6] use the arrangement strategy, part-based, chain-based, and traversal-based orders, to convert the skeleton sequences into color images. Yang et al. [82] take advantage of both CNN and RNN to capture semantic structural information and long-term dependencies. In recent years, GCN generalizes CNN to graph-structured data with graph theory. GCN can efficiently distill information from skeleton sequences. Specifically, Shi et al. [61] propose a two-stream GCN to model both first-order and second-order information adaptively. Li et al. [39] build AS-GCN by stacking actional structural graph convolution and temporal convolution to model both spatial and temporal features.

### 2.2.2 FPV methods

Moghimi et al. [50] design a domain-specific feature that roughly encodes the distribution and location of skin pixels in the image. The skin segmentation is the first step of the proposed method which includes both RGB and depth information. Then, a variety of descriptors, such as skin histogram and arms-hands bounding box, are built on the top of the obtained skin pixel segmentation and are sent into the classifier for action recognition. Xia et al. [79] select four different descriptors that have shown to perform well in classic activity recognition tasks: 3D optical flow, spatiotemporal interesting points in RGB data, depth data, and body posture descriptors. Moreover, they generate independent motion vectors and attention masks to separate ego-motion from independent motion. Tang et al. [65] develop a multi-stream deep neural network (MDNN) to adequately learn the nonlinear structure of heterogeneous representations from RGB and depth modality as well as exploit the complementary characteristics in TPV videos. It can preserve the distinctive property for each modality and simultaneously explore sharable information in a unified deep architecture. Moreover, in some research, the methods for TPV action recognition, such as two-stream network [14] and VGG16 [62] for RGB videos and HON4D [53] for depth maps are directly used in RGB-D FPV action recognition.

### 3 The FT-HID dataset

This section presents the details and the evaluation protocols of the collected FT-HID dataset. Figure 1 shows some sample frames of the dataset.

#### 3.1 Action classes

FT-HID dataset contained 30 interaction categories for both first- and third-person vision in total. A summary of the quantitative characteristics of the dataset is presented in Table 2. The interactions can be divided into two groups, including 23 positive interactions and seven negative actions. In particular, positive interactions commonly happen in daily life, while negative actions are abnormal actions that usually tend toward violence. The overall categories are as follows.

##### 3.1.1 Positive interactions

- Push hands, Cross-clap hands, Clap hands, Wave hands, Bow, Shake hands, High-five, Fist to fist, Finger-guessing, Celebration, Thumbs up, Pour water, Cheers, Pass ball by hand, Pass ball by feet, Pass an object, Feet-guessing, Lift the table, Wrist wrestling, Shuffle the cards, Take cards, Take out cards, and Touch the pulse.

##### 3.1.2 Negative actions

- Beat hands, Parry, Pull, Kick, Push, Steal from the back pocket and Collide with each other.

#### 3.2 Subjects

One hundred and nine distinct individuals participated in the dataset collection phase. They chose their partner freely and finally formed 89 two-person interaction groups with consistent ID numbers over the entire dataset. For every action category, each group performed four times and the FPV camera wearer changed in action interval.

#### 3.3 Scenarios

We provided two different indoor backgrounds with various complex objects, such as desk, armchair, computer,
and air conditioner. Three fixed cameras were used to capture TPV videos and the filming range was settled. The participants could walk freely within the range. We encouraged the participants to perform as naturally as possible. Thus, the category name of each action was the only hint, and after hearing the category, participants started to perform those actions according to their intuition.

3.4 Camera settings and modalities

3.4.1 First-person vision

Considering the available modality, weight, and volume, Orbbec Persee that provided both RGB videos and depth information was utilized in the FPV video capture. The camera was mounted on the helmet and the wearer can move effortlessly as usual. The depth maps and RGB videos were collected and stored about 25 frames per second at the resolution of 320 × 240. Since the videos for each interaction group were recorded continuously, eight participants were invited to assist video segmentation. They revisited RGB videos and pointed out the exact starting and ending frames of each action. As a result, the original 321 videos were segmented into 8,718 samples.

3.4.2 Third-person vision

While collecting FPV videos, data from TPV were recorded synchronously. We used three Kinect v2 to capture two different horizontal views and a vertical view for the same action. The front and side cameras were set up at the same distance, 1.8 meters away from the performance area, yet from two different horizontal angles: 0° and 45°. The top camera was about 2.4 meters from the ground. We collected three major data modalities acquired by the deep camera, namely the RGB videos, the depth maps, and the 3D skeleton sequences. But the skeleton information was defective in the top camera, thus we only provided RGB videos and depth maps for the top view. The RGB videos were recorded in the provided resolution of 1920 × 1080 and contained 9,882 samples per view. The resolution of each depth map was 512 × 424. The skeleton sequence consisted of 3D coordinates of 25 major body joints for each tracked human body in the scene.

3.5 Evaluation protocols

In order to fairly evaluate the performances of different methods applied on this dataset, we define several precise evaluation criteria for action recognition. For each criterion, we report the classification accuracy in percentage.

3.5.1 Cross-subject protocol

In cross-subject (CS) evaluation, 89 interaction groups are split into training and testing groups. For this evaluation, we state two sample split formulation. In the First...
4.1 Multi-view interaction network for skeleton-based action recognition

To deal with the data from multiple views, some efforts have been made over the past years. Some works [41, 87] unify the data from different views into a certain coordinate system through rotor and matrix transformation, but the view-specific information might be lost in the process of transformation. Some other works [59, 73] utilize parallel networks to extract the features of different views independently, which neglect the correlations among views. However, the view-specific information is beneficial to tackle the occlusions, and the relationships among views can improve the tracking quality. Therefore, considering the distinctions and complementarity of different views, we propose a selective interaction network for multi-view skeleton-based action recognition.

4 Baseline methods

In this section, we introduce two new data-driven learning methods to model human actions using the collected 3D action sequences. Specifically, one focuses on the channel correlations in deep models inspired by the current attention mechanism. The other is for paired first- and third-person vision, which can make full use of the advantages of both perspectives.
\[ d^l = GDC(l^l), d^l \in R^{1 \times 1 \times MC^l} \]

Further, two fully connected (FC) layers are employed to obtain the correlation representation of all channels:

\[ r^l = FC_2(FC_1(r^l)), r^l \in R^{1 \times 1 \times MC^l} \]

Note that the output channels of the second FC layer must remain the same as \( d^l \) for the subsequent selection. Here, \( r^l \) can be seen as the stack of \( M \) vectors which are relative to the features before concatenation:

\[ r^l = \{r^l_1, ..., r^l_M\}, r^l_m \in R^{1 \times 1 \times C^l} \]

In order to enhance the informative views and weaken the other ones, \( \text{softmax} \) function is used across description vectors of different views. The resulting weight vectors \( w^l \), whose value is in the range of 0 to 1, represent the importance of different views in certain channels:

\[ w^l = \{w^l_1, ..., w^l_M\} = \text{softmax}\{r^l_1, ..., r^l_M\} \]

where \( w^l \in R^{1 \times 1 \times MC^l} \). The final feature map \( P^l \) is obtained through the weight vectors on various views and the residual connection with the previous feature map:

\[ P^l = w^l \cdot A^l + A^l \]

\[ = \{w^l_1 \cdot A^l_1 + A^l_1, ..., w^l_M \cdot A^l_M + A^l_M\} \]

\[ = \{P^l_1, ..., P^l_M\}, P^l_m \in R^{F \times F \times C^l} \]

Then, the enhanced feature maps \( P^l_m \) of each view are sent into the next convolutional layer for higher-level feature extraction.

The proposed MVIB optimizes the network representation from two aspects. First, for a single view, the weights of different channels indicate their significance for recognizing the action, which achieves channel selection. Second, for multiple views, each view obtains a corresponding importance weight, which contains the correlation with other views, for a certain channel. And the accumulation of these weights along channels demonstrates the contribution of each view to the final classification. Therefore, the proposed mechanism can not only emphasize the vital channels within the interlayer features, but also can learn the interaction information between multiple views and guide the network to focus on informative views.

4.2 Dynamic sequences for FPV and TPV learning

In current FPV action recognition methods, video frames are generally directly sent into the neural networks for feature extraction. This ignores the redundant information in the sequence and fails to highlight the distinctive motion patterns. While for TPV action recognition, some works \([10, 15, 91]\) compress the initial data through rank pooling to model the video-wide temporal evolution. However, most existing rank pooling-based methods generate a
single dynamic image for an entire action video, which may lose important sub-action dynamics. By contrast, the proposed method extracts sub-segments using the pre-defined temporal filters and generates multiple dynamic images corresponding to these sub-segments. Moreover, in order to capture both short-term and long-term spatiotemporal features of the compressed images, we propose SI3D-ConvLSTM Network. The generation of sub-segment dynamics and detailed network architecture is as follows.

### 4.2.1 Sub-segment dynamics

Our main motivation is to exploit the changing of the behavior of the appearance in the frames over time, hence revealing the motion dynamics in short intervals. This is done by applying a handcrafted temporal filter and a rank pooling mechanism to capture temporal statics and features.

Concretely, assume a set of $N$ action videos, $V = \{V_1, ... , V_n, ... , V_N\}$, a video instance $V_n$ with $F$ frames can be represented by a set of ordered continuous frames, $V_n = \langle v_1^n, ... , v_r^n, ... , v_F^n \rangle$, $v_{ \cdot }^n \in \mathbb{R}^d$. From these sequences of vectors, we adaptively extract the sub-segments using a defined window size $w$ and stride $s$ for $t_r = \{1, r+1, 2r+1, ... \}$, where $t_r$ denotes the starting frame index of the local time interval, i.e., $\{[t_r, t_r + 1), ... , [t_r^F, t_r^F]\}$. Then, the resultant sub-segments $G = \{g_{\cdot}^k\}_{k=1}^H$ are sent into rank pooling module to get dynamic representations. In each sequence, the number of sub-segments $H$ is fixed to fit the network thus the window size $w$ and stride $s$ varies with the sequence lengths as follows:

\[
\begin{align*}
&\begin{cases} 
  w = F - (H - 1), F < \varepsilon \\
  w = \text{floor}(F / H), F \geq \varepsilon 
\end{cases} \\
&\begin{cases} 
  s = 1, F < \varepsilon \\
  s = w, F \geq \varepsilon 
\end{cases}
\end{align*}
\]

(7) \quad (8)

where $\varepsilon$ is a predefined threshold that divides long and short action sequences. The temporal order has shown to be an important feature for video representation as it can effectively model the evolution of the frame appearance. Therefore, we propose to encode the sub-segment dynamics by employing a rank pooling algorithm on the segments obtained from the temporal filter. Specifically, we are interested in getting a ranking machine to model a regression space that explicitly imposes a strict criterion that earlier frames must precede the following frames. If $g_{\cdot}^{b+1}$ succeeds $g_{\cdot}^b$, we have an ordering denoted by $g_{\cdot}^{b+1} \succ g_{\cdot}^b$. To encode the dynamics of a sub-segment, the ranking function $\psi_{\cdot}(g_{\cdot}^b; u) = u^T \cdot g_{\cdot}^b$ parametrized by vector $u$. As a result, the aim is to learn the parametric vector $u$ such that it satisfies all constraints $\forall g_{\cdot}^a, g_{\cdot}^b, g_{\cdot}^c \succ g_{\cdot}^d \iff u^T \cdot g_{\cdot}^a > u^T \cdot g_{\cdot}^b$. Using the principle of RankSVM [29] for structural risk minimization and max-margin formulation, the objective function can be expressed as:

\[
\begin{align*}
\arg\min_u & \frac{1}{2} \|u\|^2 + C \sum_{ab, g^a \succ g^b} \varepsilon_{ab} \\
\text{s.t.} & \quad u^T \cdot (g^a - g^b) \geq 1 - \varepsilon_{ab} \\
& \quad \varepsilon_{ab} \geq 0
\end{align*}
\]

(9)

To estimate this linear function and learn the parametric vector $u$, we use the support vector regressor (SVR) for its efficiency. The learned vector $u$ defines the pooled feature representation for each sub-segment.

### 4.2.2 Global spatiotemporal extraction and classification

After the sub-segment dynamics have been encoded, neural networks are adopted to model the global spatiotemporal information of the entire video. ConvLSTM is first proposed in [80], which can not only establish temporal relations like LSTM, but also depict local spatial features like CNN.

Many experiments proved that ConvLSTM is more effective than LSTM and CNN in obtaining spatiotemporal relations. Therefore, we employ ConvLSTM on the resultant dynamic maps for potent feature extraction. Moreover, to reduce parameters of the network and improve computation efficiency, we design a 3D feature extractor to reduce the dimension of the input sequence and extract short-term spatiotemporal features before three ConvLSTM layers. Specifically, the 3D feature extractor consists of several 3D convolution layers, inception modules [7] and pooling layers. Finally, the learned high-level representations are fed into a global average pooling layer and two fully connected layers followed by a softmax activation function for action classification. The entire framework named SI3D-ConvLSTM Network is shown in Fig. 3.

### 4.2.3 TPV and FPV Joint learning

In the FT-HID dataset, TPV consists of three different views while FPV is taken from a single view. Both of them contain RGB videos and depth maps. All kinds of initial data are firstly sent into the sub-segment dynamics generation module for short-term temporal relations modeling.

Since TPV and FPV are recorded synchronously and follow the same action instance annotation strategy, the dynamic maps for different inputs are one-one correspondence. The dynamics are sent into separate networks for independent global spatiotemporal feature extraction. Score fusion mechanisms are used to combine the
predictions of the total eight streams and produce the final recognition category of the current action. The fusion process can take advantage of all kinds of data and enhance the performance of action recognition.

5 Experiments

In this section, some conventional handcrafted methods and deep learning methods are evaluated based on the suggested cross-view and cross-subject criteria. Then, the performance of the two proposed networks for skeleton-based action recognition and paired-vision action recognition are also evaluated, respectively.

5.1 Results of existing action recognition methods

We use the publicly available implementation of eight action recognition approaches and apply them to our new dataset. Among them, HOPC [55], HON4D [53], DMMLBP [8], DMMCNN [72], and RankPool [15] extract features directly from depth maps without using the skeleton information. JDM [37], JTM [73] and IDSF [11] are skeleton-based methods, for which the skeleton data from the front view are used for evaluation. Except the skeleton-based methods [11, 37, 73], both CS and CV evaluation criteria are used on all the other approaches [8, 15, 53, 55, 72], and the results are reported in Tables 3, 4, and Fig. 4.

As seen, in both CS and CV evaluation, deep learning-based methods [15, 72] achieve better results than handcrafted-based methods [8, 53, 55]. In Table 3, all the results of using the second formulation of CS evaluation are lower than results using the first formulation. It indicates that the second formulation has larger data variation between the training set and the testing set, and is more challenging for existing action recognition approaches. Besides, it can be seen that the result of the front view is the best in most situations because the video captured from the front of the subject can catch more distinctive information than other views. Moreover, the performance of the top view is better than the side view. It reflects the importance of the top view data, which is always be omitted in other action recognition datasets. For some of the methods [8, 55], the recognition results of FPV are comparable to TPV. While for other methods [53, 72], they are apparently worse than TPV. This reveals that the performance of many existing methods is unsatisfactory for FPV action recognition, and it is urgent to explore methods that can accommodate TPV and FPV simultaneously. The experimental results of the skeleton-based methods under the CS evaluation criteria are summarized in Table 4. For the first formulation, it can be observed that the three methods outperform the result of the second formulation by 2.98, 7.93, and 5.10 percentage points, respectively. It further indicates the difficulty of the second formulation. Note that the three skeleton-based methods perform better than all the depth-based methods, which verify the efficiency of the skeleton data. JTM [73] encodes skeleton joint trajectories of each time instance into HSV images, while JDM [37] employs joint distance.
as spatial feature with color encoding. IDSF [11] adopts five skeletal features, including joint-joint distance, joint-joint orientation, joint-joint vector, joint-line distance, and line-line angle. Among the three methods, JDM achieves the best recognition accuracy in both formulations and increases the evaluation result by 9.71 and 14.66 percentage points, respectively, compared with JTM. The results of the CV evaluation criteria are shown in Fig. 4. HOPC [55], HON4D [53], and RankPool [15] are handcrafted methods, while DMMCNN [72] is a deep learning-based method that comprises the weighted hierarchical depth motion maps and a three-channel convolutional neural network. As shown in Fig. 4, the deep learning-based DMMCNN increases the accuracy of HOPC by 46.23 percentage points and also outperforms other handcrafted methods by a large margin, which indicates the effectiveness of neural networks.

Under the CS evaluation criteria, four score fusion strategies, including addition, multiplication, maximum, and a so-called one channel fusion strategy are used to integrate the front, side, and top views in TPV. For the first three score fusion methods, data from different views are trained by three independent networks and the score of each category is fused after the fully connected layer. For the one channel fusion strategy, data from three views are sent them to one network for training. The fusion results of the three views in TPV are reported in Fig. 5. It is clear that the recognition accuracy raises a lot after combining the three camera views but is still restricted by the capability of methods. Meanwhile, the one channel method is the worst among the four fusion strategies. The reason is that the data from the three views are quite different, especially the front and top views. It is difficult for the network to learn effective features when they are treated equally without discrimination, while the separate training can maintain the characteristics of the complementary views and the score fusion improves the performance. Among the four fusion strategies, the multiplicative score fusion in the first formulation consistently outperforms the addition, max score

### Table 3: Results of different methods, which are designed for depth-based human action analysis, using the two formulations of the CS evaluation criteria on our constructed FT-HID dataset

| Method  | Front view | Side view | Top view | First view | Front view | Side view | Top view | First view |
|---------|------------|-----------|----------|------------|------------|-----------|----------|------------|
| HOPC [55] | 0.5132     | 0.4971    | 0.5559   | 0.4798     | 0.4632     | 0.4420    | 0.5007   | 0.4501     |
| HON4D [53] | 0.6135     | 0.1864    | 0.5988   | 0.1957     | 0.5903     | 0.0664    | 0.4614   | 0.1901     |
| DMMNB [8]  | 0.5979     | 0.5885    | 0.6023   | 0.5550     | 0.1484     | 0.4980    | 0.5644   | 0.5158     |
| DMMCNN [72] | 0.7953     | 0.6070    | 0.6654   | 0.2904     | 0.7122     | 0.2331    | 0.5513   | 0.3660     |
| RankPool [15] | 0.7474     | 0.5037    | 0.6413   | 0.4936     | 0.5720     | 0.3370    | 0.6310   | 0.4503     |

### Table 4: Results of different methods, which are designed for skeleton-based human action analysis, using the two formulations of the CS evaluation criteria on the FT-HID dataset

| Method  | First formulation | Second formulation |
|---------|-------------------|--------------------|
| JDM [37]  | 0.8603            | 0.8305             |
| JTM [73]  | 0.7632            | 0.6839             |
| IDSF [11] | 0.7788            | 0.7278             |

Fig. 4 Performance of different methods on the FT-HID dataset using the CV evaluation criteria

Fig. 5 Comparison of the fusion strategy of TPV with two CS settings. F1 and F2 refer to the first formulation and the second formulation, respectively. OC refers to the one channel fusion strategy.
fusion, and one channel methods. This verifies that the feature of the three views is likely to be statistically independent and provide complementary information. However, for some methods, the multiplicative score fusion performs worse than other fusion strategies in the second formulation. It indicates that the design of feature fusion strategy should consider the characteristics and distribution of data.

In order to verify the complementarity between TPV and FPV, the fusion strategy is also applied to these two visions. Considering the characteristics of TPV and FPV are quite different, they are trained separately. Specifically, for the different views contained in TPV, the aforementioned one-channel scheme is used for holistic feature extraction. Then, combine TPV and FPV by the score fusion strategy. The results are reported in Fig. 6. From this, we can see that on the evaluated five methods, the integration of two distinct visions significantly improves the performance of the single vision.

5.2 Results of the multi-view interaction network on skeletons

We evaluate the performance of the proposed multi-view interaction network on our dataset. Since the FT-HID dataset provides the skeleton data from front and side view, the network in this experiment is a two-stream architecture. The data partition mechanism follows the CS evaluation criteria, and both the first and the second formulations are taken into account. To further verify the efficiency of the proposed MVIB module, two contrasting multi-view fusion strategy is designed. In the first contrasting fusion strategy, all the MVIB are removed from the proposed multi-view interaction network, and the final concatenation operation is retained. The second contrasting fusion strategy is to train two independent networks and then adopt multiply score fusion. The results are shown in Table 5. It can be seen that the concatenate fusion performs better than the score fusion, which indicates that using an entire network can better utilize the complementarities among multiple views of skeleton data. Moreover, it is clear that the proposed MVIB module outperforms the other two fusion strategy in both settings, which shows the superiority of the MVIB in multi-view data interaction.

The confusion matrix of the proposed network with the second CS formulation is shown in Fig. 7. To perform in-depth action-wise analysis, action classes that have high recognition accuracy (top five accurate classes), and the actions that are easily misclassified to other classes (top five confused action pairs) are listed in Table 6. Based on the results, it can be found that the actions that have significant motions and discriminative posture patterns could be more accurately recognized. For example, the motion-rich actions “Bow” and “cheers” are in the top five accurate actions. Moreover, it can be observed that for skeleton data, the actions that involve interactions with objects may be easily misclassified. For example, “Shuffle the cards” and “Pass ball by feet” are often confused with “Lift the table,” and “Pass an object” is misclassified to “Celebration.” This is possible because the misclassified action pairs have similar human motion patterns, and the perception of the existences of the objects and their appearances is important for accurately recognizing these actions, which cannot be learned when using skeleton information only.
5.3 Results of dynamic mapping and SI3D-ConvLSTM network for TPV and FPV learning

We evaluate the performance of the dynamic mapping and the designed SI3D-ConvLSTM network on the FT-HID dataset. In this experiment, we utilize the TPV and FPV combined protocol. Specifically, the RGB and depth modality of the front view of TPV and FPV are taken into account. The results are shown in Fig. 8. To verify the effectiveness of the proposed SI3D-ConvLSTM network, two 3D-ConvLSTM networks, I3D-ConvLSTM and Conv3D-CLSTM [90], are also evaluated on the FT-HID dataset. For the fairness of comparison, three networks have the same ConvLSTM layers and fully connected layers, while the spatial pyramid pooling layer in Conv3D-CLSTM is removed. The first four groups in Fig. 8 show the results of the single view and single modality data, including TPV RGB (T-RGB) data, TPV Depth (T-Depth) data, FPV RGB (F-RGB) data, and FPV Depth (F-Depth) data. The intermediate four groups depict the fusion result of each two of the four single inputs, for example, the RGB is the combination of T-RGB data and F-RGB data. The last group illustrates the fusion of all the four single inputs (TFRD). It is clear that in both single input and fusion conditions, the proposed SI3D-CLSTM achieves the highest accuracy and shows its superiority. The confusion matrix of the proposed network in TFRD fusion condition with the second formulation of CS evaluation criteria is shown in Fig. 9.

Similar to the analysis of the multi-view interaction network, the top five accurate actions and the top five confused action pairs of five fusion conditions are listed in Table 7. It can be observed that in contrast to skeleton data, both RGB and depth data perform well in capturing the object information, many object-related actions could be accurately recognized. For example, “Pass ball by feet” and “Pass ball by hand” are in the top five accurate actions of five conditions. Although RGB and depth data modalities both have a good ability in representing the object-related actions, the actions with the object involved may be misclassified when the same object and similar motion

| Top 5 accurate | Top 5 confused pairs |
|----------------|----------------------|
| Lift the table | Beat hands—Lift the table |
| Cheers         | Pass an object—Celebration |
| Bow            | Shuffe the cards—Lift the table |
| Take out cards | Push—Steal from the back pocket |
| Wrist wrestling| Pass ball by feet—Lift the table |

Fig. 9 Confusion matrix of the SI3D-convLSTM Network with the second CS formulation
patterns are shared by different actions. For example, “Take cards” tend to be misclassified to “Take off cards.” Compared with the TPV data, FPV data suffers more occlusion which makes it easily misclassified to some actions that have similar local motion. For example, when using FPV data as input, “Parry” and “Push” are confused, “Pass an object” tends to be misclassified to “Celebration.” As a result, the TPV data, FPV data, RGB appearance data, and depth data can all affect the recognition performance of different types of action. Meanwhile, the rational fusion of different views and modalities which provide complementary appearance or geometric information may improve the recognition accuracy.

6 Conclusion

A large-scale RGB-D action recognition dataset with pair-aligned TPV and FPV information is introduced in this paper. The dataset has been validated by existing action recognition methods and also by two extended methods. Compared to the existing datasets for the action recognition, the proposed FT-HID dataset has unique features including a large number of samples, distinct interaction categories, diverse camera views, various environments, and a wide range of participants. The FT-HID dataset is expected to facilitate further research in TPV- and FPV-based human action recognition using individual modalities, multi-modalities, and cross-modalities protocols that covers a broad range of real applications.

Acknowledgements The work of Zhimin Gao was supported in part by National Natural Science Foundation of China under Grant No. 61906173.

Data availability The FT-HID dataset that support the findings of this study is available in Baidu Cloud, https://pan.baidu.com/s/1RHMRF-O8VLljLo5j9DxoRA?pwd=wr6u. All other data are available from the authors upon reasonable request.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Asadi-Aghbolaghi M, Bertiche H, Roig V, Kasaei S, Escalera S (2017) Action recognition from rgb-d data: Comparison and fusion of spatio-temporal handcrafted features and deep strategies. In: Proceedings of the IEEE International conference on computer vision workshops, pp. 3179–3188
2. Ben Tanfous A, Drira H, Ben Amor B (2018) Coding kendall’s shape trajectories for 3d action recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2840–2849
3. Bloom V, Argyriou V, Makris D (2013) Dynamic feature selection for online action recognition. In: international workshop on human behavior understanding, pp. 64–76. Springer
4. Bloom V, Argyriou V, Makris D (2014) G3di: A gaming interaction dataset with a real time detection and evaluation framework. In: European conference on computer vision, pp. 698–712. Springer
5. Bloom V, Makris D, Argyriou V (2012) G3d: a gaming action dataset and real time action recognition evaluation framework. In: 2012 IEEE Computer society conference on computer vision and pattern recognition workshops, pp. 7–12. IEEE
6. Cao C, Lan C, Zhang Y, Zeng W, Lu H, Zhang Y (2018) Skeleton-based action recognition with gated convolutional neural networks. IEEE Trans Circ Syst Video Tech 29(11):3247–3257
7. Carreira J, Zisserman A (2017) Quo vadis, action recognition? a new model and the kinetics dataset. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 6299–6308
8. Chen C, Jafari R, Kehtarnavaz N (2015) Action recognition from depth sequences using depth motion maps-based local binary patterns. In: 2015 IEEE winter conference on applications of computer vision, pp. 1092–1099. IEEE
9. Chen C, Jafari R, Kehtarnavaz N (2015) Utd-mhad: A multi-modal dataset for human action recognition utilizing a depth
camera and a wearable inertial sensor. In: 2015 IEEE International conference on image processing (ICIP), pp. 168–172. IEEE
10. Cherian A, Fernando B, Harandi M, Gould S (2017) Generalized rank pooling for activity recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3222–3231
11. Ding Z, Wang P, Ogumbona PO, Li W (2017) Investigation of different skeleton features for cmn-based 3d action recognition. In: 2017 IEEE International conference on multimedia & expo workshops (ICM EW), pp. 617–622. IEEE
12. Fan Z, Zhao X, Lin T, Su H (2018) Attention-based multiview re-observation fusion network for skeletal action recognition. IEEE Trans Multim 21(2):363–374
13. Fathi A, Ren X, Rehg JM (2011) Learning to recognize objects in egocentric activities. In: CVPR 2011, pp. 3281–3288. IEEE
14. Feichtenhofer C, Pinz A, Zisserman A (2016) Convolutional two-stream network fusion for video action recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1933–1941
15. Fernando B, Gavves E, Oramas J, Ghodrati A, Tuytelaars T (2016) Rank pooling for action recognition. IEEE Trans Patt Anal Mach Intell 39(4):773–787
16. Gao X, Hu W, Tang J, Liu J, Guo Z (2019) Optimized skeleton-based action recognition via sparsified graph regression. In: Proceedings of the 27th ACM international conference on multimedia, pp. 601–610. ACM
17. Gao Z, Li S, Zhu Y, Wang C, Zhang H (2017) Collaborative sparse representation learning model for rgb-d action recognition. J Visual Commun Image Represent 48:442–452
18. Gao Z, Zhang H, Xu G, Xue Y (2015) Multi-perspective and multi-modality joint representation and recognition model for 3d action recognition. Neurocomputing 151:554–564
19. Garcia NC, Morerio P, Murino V (2018) Modality distillation with multiple stream networks for action recognition. In: Proceedings of the european conference on computer vision (ECCV), pp. 103–118
20. Garcia-Hernando G, Yuan S, Baek S, Kim TK (2018) First-person hand action benchmark with rgb-d videos and 3d hand pose annotations. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 409–419
21. Hou Y, Yu H, Zhou D, Wang P, Ge H, Zhang J, Zhang Q (2021) Local-aware spatio-temporal attention network with multi-stage feature fusion for human action recognition. Neur Comp Appl 33(23):16439–16450
22. Hu JF, Zheng WS, Lai J, Zhang J (2015) Jointly learning heterogeneous features for rgb-d action recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 5344–5352
23. Hu JF, Zheng WS, Pan J, Lai J, Zhang J (2018) Deep bilinear learning for rgb-d action recognition. In: Proceedings of the European conference on computer vision (ECCV), pp. 335–351
24. Huang D, Yao S, Wang Y, De La Torre, F (2014) Sequential max-margin event detectors. In: European conference on computer vision, pp. 410–424. Springer
25. Ilijina EP, Chalavadi KM (2017) Human action recognition in rgb-d videos using motion sequence information and deep learning. Patt Recognit 72:504–516
26. Imran J, Kumar P (2016) Human action recognition using rgb-d sensor and deep convolutional neural networks. In: 2016 International conference on advances in computing, communications and informatics (ICACCI), pp. 144–148. IEEE
27. Ji Y, Xu F, Yang Y, Shen F, Shen HT, Zheng WS (2018) A large-scale rgb-d database for arbitrary-view human action recognition. In: 2018 ACM Multimedia conference on multimedia conference, pp. 1510–1518. ACM
28. Jia C, Fu Y (2016) Low-rank tensor subspace learning for rgb-d action recognition. IEEE Trans Image Process 25(10):4641–4652
29. Joachims T (2006) Training linear svms in linear time. In: Proceedings of the 12th ACM SIGKDD International conference on Knowledge discovery and data mining, pp. 217–226
30. Khowaja SA, Lee SL (2020) Hybrid and hierarchical fusion networks: a deep cross-modal learning architecture for action recognition. Neur Comp Appl 32(14):10423–10434
31. Kong J, Liu T, Jiang M (2019) Collaborative multimodal feature learning for rgb-d action recognition. J Visu Commun Image Represent 59:537–549
32. Kong Y, Fu Y (2015) Bilinear heterogeneous information machine for rgb-d action recognition. In: Proceedings of the IEEE Conference on computer vision and pattern recognition, pp. 1054–1062
33. Kong Y, Fu Y (2016) Discriminative relational representation learning for rgb-d action recognition. IEEE Trans Image Process 25(6):2856–2865
34. Kong Y, Fu Y (2017) Max-margin heterogeneous information machine for rgb-d action recognition. Int J Comp Vision 123(3):350–371
35. Koperski M, Bremond, F (2016) Modeling spatial layout of features for real world scenario rgb-d action recognition. In: 2016 13th IEEE international conference on advanced video and signal based surveillance (AVSSS), pp. 44–50. IEEE
36. Li B, Li X, Zhang Z, Wu F (2019) Spatio-temporal graph routing for skeleton-based action recognition. In Proceedings of the AAAI Conference on Artificial Intelligence 33(1), pp 8561-8568
37. Li C, Hou Y, Wang P, Li W (2017) Joint distance maps based action recognition with convolutional neural networks. IEEE Sign Process Lett 24(5):624–628
38. Li C, Li S, Gao Y, Zhang X, Li W (2021) A two-stream neural network for pose-based hand gesture recognition. IEEE Trans Cognit Develop Sys. https://doi.org/10.1109/TCDS.2021.3126637
39. Li M, Chen S, Chen X, Zhang Y, Wang Y, Tian, Q (2019) Actional-structural graph convolutional networks for skeleton-based action recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3595–3603
40. Li W, Zhang Z, Liu Z (2010) Action recognition based on a bag of 3d points. In: 2010 IEEE computer society conference on computer vision and pattern recognition-workshops, pp. 9–14. IEEE
41. Li Y, Xia R, Liu X (2020) Learning shape and motion representations for view invariant skeleton-based action recognition. Patt Recognit 103:107293
42. Liu AA, Nie WZ, Su YT, Ma L, Hao T, Yang ZX (2015) Coupled hidden conditional random fields for rgb-d human action recognition. Sig Process 112:74–82
43. Liu AA, Xu N, Nie WZ, Su YT, Hong Y, Kankanhalli M (2016) Benchmarking a multimodal and multiview and interactive dataset for human action recognition. IEEE Trans Cybern 47(7):1781–1794
44. Liu AA, Xu N, Su YT, Lin H, Hao T, Yang ZX (2015) Single/multi-view human action recognition via regularized multi-task learning. Neurocomputing 151:544–553
45. Liu H, Yuan M, Sun F (2015) Rgb-d action recognition using linear coding. Neurocomputing 149:79–85
46. Liu J, Akhtar N, Ajmal M (2018) Viewpoint invariant action recognition using rgb-d videos. IEEE Access 6:70061–70071
47. Liu J, Shahroudy A, Perez ML, Wang G, Duan LY, Chichung AK (2019) Ntu rgb+d 120: a large-scale benchmark for 3d human activity understanding. IEEE Trans Patt Anal Mach Intell 42(10):2684–2701
85. Yun K, Honorio J, Chattopadhyay D, Berg TL, Samaras D (2012) Two-person interaction detection using body-pose features and multiple instance learning. In: 2012 IEEE Computer society conference on computer vision and pattern recognition workshops, pp. 28–35. IEEE
86. Zhang C, Tian Y (2012) Rgb-d camera-based daily living activity recognition. J Comp Vis Image Process 2(4):12
87. Zhang P, Lan C, Xing J, Zeng W, Xue J, Zheng N (2019) View adaptive neural networks for high performance skeleton-based human action recognition. IEEE Trans Patt Anal Mach Intell 41(8):1963–1978
88. Zhang S, Yang Y, Xiao J, Liu X, Yang Y, Xie D, Zhuang Y (2018) Fusing geometric features for skeleton-based action recognition using multilayer lstm networks. IEEE Trans Multim 20(9):2330–2343
89. Zhang Y, Cao C, Cheng J, Lu H (2018) Egogesture: a new dataset and benchmark for egocentric hand gesture recognition. IEEE Trans Multim 20(5):1038–1050
90. Zhu G, Zhang L, Shen P, Song J (2017) Multimodal gesture recognition using 3-d convolution and convolutional lstm. IEEE Access 5:4517–4524
91. Zhu Z, Ji H, Zhang W, Xu Y (2018) Rank pooling dynamic network: Learning end-to-end dynamic characteristic for action recognition. Neurocomputing 317:101–109
92. Zong M, Wang R, Chen Z, Wang M, Wang X, Potgieter J (2021) Multi-cue based 3d residual network for action recognition. Neur Comp Appl 33(10):5167–5181

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.