Front. Comput. Sci., 2024, 18(5): 185325
https://doi.org/10.1007/s11704-023-2418-0

RESEARCH ARTICLE

Contactless interaction recognition and interactor detection in multi-person scenes

Jiacheng LI[1], Ruize HAN (✉)[1], Wei FENG[1], Haomin YAN[1], Song WANG[2]

1 College of Intelligence and Computing, Tianjin University, Tianjin 300350, China
2 Department of Computer Science and Engineering, University of South Carolina, Columbia SC 29208, USA

Abstract Human interaction recognition is an essential task in video surveillance. The current works on human interaction recognition mainly focus on the scenarios only containing the close-contact interactive subjects without other people. In this paper, we handle more practical but more challenging scenarios where interactive subjects are contactless and other subjects not involved in the interactions of interest are also present in the scene. To address this problem, we propose an Interactive Relation Embedding Network (IRE-Net) to simultaneously identify the subjects involved in the interaction and recognize their interaction category. As a new problem, we also build a new dataset with annotations and metrics for performance evaluation. Experimental results on this dataset show significant improvements of the proposed method when compared with current methods developed for human interaction recognition and group activity recognition.

Keywords human-human interaction recognition, multiperson scene, contactless interaction, human relation modeling

1 Introduction

Human-human interaction (HHI) recognition is an essential task in social scene understanding, which has many applications in video content analysis, such as social relation analysis [1], pedestrian trajectory tracking and prediction [2,3], and abnormal behavior analysis [4,5]. In recent years, it has been attracting more interests, which is evidenced by the release of many HHI datasets, such as UT interaction [6], SBU Kinetic Interaction [7], and AVA [8] datasets. These datasets contain different categories of human interactions, such as hugging, shaking and patting, and significantly boost the research on human interaction recognition.

There are two common characteristics for most of the existing HHI datasets, as well as the human recognition methods developed/evaluated on these datasets: 1) The subjects involved in an interaction are usually *close to each other*, e.g., two persons in hand-shaking have physical contact with each other. 2) Except for the interactive subjects, there are no other subjects in the scene, or the interactive subjects obviously *dominate the content of the image*, as shown in Figs. 1(a) and 1(b). However, these data and the algorithms developed on these data do not reflect the complexity in practice. In many real scenes, the interactive activities occur in the multi-person scene, e.g., a party or other social events, which usually include *many other subjects not involved in the interaction*, as shown in Fig. 1(d). Further, in many applications, we are interested in the interactions where the involved subjects may not have body contact and even keep a distance from each other. For example, two persons may greet each other by *keeping a social distance* (e.g., under the situation with pandemic), and such an interaction is contactless. As shown in Figs. 1(c) and 1(d), in this work, we aim to study a new problem of **contactless interactive activity recognition in the multi-person scenes**, i.e., we both identify the interactive subjects from non-involved subjects and recognize the interaction category.

This work takes a step to extend the current HHI research to HHI in more practical multi-person scene, which has many potential applications and could benefit the real-world video surveillance and abnormal behavior analysis [9−11]. As a new task, in this work, we first focus on a basic and common situation, where one interaction of interest involving two

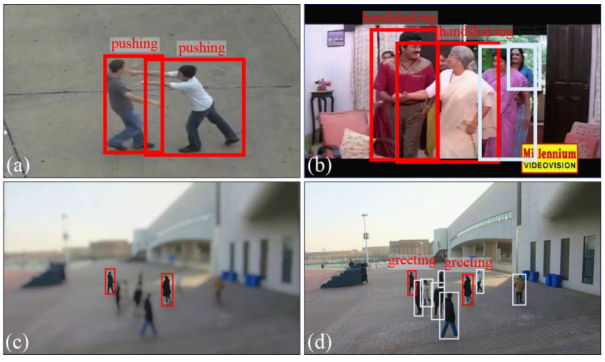


Fig. 1 An illustration of different interactive activities. (a) "Pushing" in UT interaction dataset and (b) "Shaking" in AVA dataset. (c−d) Contactless interactive activities in the multi-person scene that are studied in this paper, where red bounding boxes indicate the interactive subjects

Received July 2, 2022; accepted May 30, 2023

E-mail: han_ruize@tju.edu.cn

interactors appears in the scene. Even so, *given the coupling of contactless-interaction characteristic and multi-person scene*, this problem is much more challenging than those studied in previous works. First for the multi-person scene, without knowing the interactive subjects that are conspicuous in existing HHI [8,12,13], we have to take *the relations among all the subjects* into consideration. Second, the distance between the contactless interactors is *not distinguishing* with the distances among other subjects, thus the interactors are not easy to be discovered by simple spatial distances among the subjects like previous works [1].

Specifically, this problem is **different from those studied in the existing HHI datasets**, e.g., UT-interaction, SBU Kinetic Interaction, BIT [14], and ShakFive2 [15], in which only the interactive subjects are present in the scene without other salient subjects. While some other datasets, like AVA, contain multiple-person scenes and some interaction categories without direct contact, they are different from our task in the following aspects. First, in AVA the interactive subjects usually get close with each other and are always the focal targets of the camera due to the nature of movie shooting style. This makes the salient interactive subjects to be easily cropped/highlighted out from the scene. Besides, only a small portion of videos in these datasets contain more than two humans or the interactions without body contact. Prior research on group activity recognition also involves the scene with a group of people. But group-activity recognition methods cannot be directly applied to solve our problem since they usually take the information from *most people* in the scene to predict an *overall* group activity. While in our problem, we aim to detect the *local* human-human interaction – the combined use of all/most subjects may easily overwhelms the features of the interactive subjects.

In this work, we propose a new interactive relation embedding network (IRE-Net), which aims to simultaneously identify the interactive subjects (interactors) and recognize their interaction category. Specifically, in IRE-Net, we first apply both the appearance and spatial information of each subject for individual representation. We further design a novel pairwise-interactive-relation cube structure to represent the relations between each pair of subjects. We finally develop a multi-head multi-task module to simultaneously predict the interactive relation and interaction category in the multi-person scenes. To verify the effectiveness of our method, we collect a new dataset for contactless HHI recognition in multi-person scenes.

The main contributions of this paper are:

1. In this work, we study a new problem of contactless human interaction recognition in the multi-person scenes, which is more practical for crowd video understanding compared to existing HHI recognition.
2. We develop a new IRE-Net for the proposed problem, which combines the appearance and spatial features and further uses a novel interactive relation embedding cube to achieve the individual-to-group short-to-long feature aggregation. Our method can be used to identify the interactor and recognize their interaction category simultaneously.
3. We build a new dataset and define new metrics for performance evaluation. We will release the videos, annotations, and evaluation toolkit to the public.

2 Related work

Human-human interaction recognition is an essential step for understanding complex human social activities and plays an essential role in surveillance video analysis. Some popular datasets include UT-interaction [6], BIT [14], SBU Kinetic Interaction [7], and ShakeFive2 [15]. They provide a wide range of interaction categories but involving only two actors with interaction in the scene. Existing HHI recognition approaches are mainly based on the human appearance [16−21] or skeleton features [22−24], or both of them [25]. Methods like [16−19] use 3D ConvNets to extend 2D image models, or decompose the convolutions into separate 2D spatial and 1D temporal filters [26−28] for capturing the spatiotemporal appearance features. Recognizing human interaction from skeleton data also attracts many interests [22−24]. Besides, several works also use the trajectory information of the person in action recognition [29,30] based on hand-crafted features. The above methods mainly focus on the HHI in the two-person scene. In most recent years, the authors in [31] propose a novel framework that simultaneously considers both implicit and explicit representations of human interactions. Also, the method in [32] adopts a hybrid learning model to the spatio-temporal relationship and occlusion relationship among the people for interaction recognition. The approach [33] addresses the multi-person human interaction recognition in images instead of the videos using the keypoint based feature image analysis. For most of the above datasets and methods, there are only interactive subjects in the scene without any other salient subjects, or the salient close-contact interactive subjects can be easily cropped/highlighted out from the scene. This is unpractical in many real-world applications.

HHI recognition in multi-person scenes has also been studied recently. For example, several works collect the video data from TV shows, films, and web videos, such as TV Human Interaction [34], AVA [8], Kinetics [12,13], and HACS [35] for HHI recognition. Besides, CMU panoptic dataset [36] provides videos of a group of people in social engagement. Most of these datasets focus on the interactions where 1) the interactive subjects get close to each other [35]; 2) the interactive actors are the focused targets in the camera due to the nature of movie shooting style [8,34]. In general, it is easy to identify the interactive actors in these datasets given that 1) the interactive actors show close contact or relative small distance with each other, and 2) the interactive actors usually exhibit dominating visual saliency in the scene. However, this is not always the case in the real world. Recently, a new dataset is proposed in [37] to detect the human social group and group activity. Based on it, the authors in [38] study the human social relation representation using a self-supervised method. Further, the panoramic human activity recognition is proposed to jointly recognize the individual action, interaction and global activity in a multi-

person scene [39]. The interactions in these works are also contacted and related to the human distance. However, many important HHIs, e.g., waving to a distant person, are not apparent to identify, especially in the multi-person scene [11]. In this work, we focus on the more challenging contactless interaction in the multi-person scenes.

Also related to our work is **group activity/relation recognition** [40−43], which usually considers both the spatial-temporal individual information and the relationship among all the subjects in the scene. Several group activity recognition methods attempt to utilize positional information by leveraging some prior knowledge [44,45]. The main difference between our task and the group activity recognition is that the latter *is more concerned about the overall video-level activity* based on the actions of all or most people in the scene [46] or some key actors with obvious actions [47]. In contrast, our task needs to *distinguish the underlying interactors* by taking all the subjects in the scene as candidates. The group relation task, e.g., the social relationship detection (SRD) [48,49] and human-object interaction (HOI) [50,51] detection, are also different from our task. Specifically, SRD aims to identify the global relation of all the involved subjects in a scene, e.g., friends and colleagues, via many human attributes, e.g., age and job. HOI is a very popular topic with many literatures in recent years [52−56]. However, HOI task depends on inherent priors of common sense and makes the interaction type highly related to the involved subjects. Differently, our task has *no such attribute or prior*, and each subject may join any type of interaction with any other subject.

3 Proposed method

3.1 Overview

We propose an Interactive Relation Embedding Network (IRE-Net) to simultaneously identify the contactless interactive subjects and recognize their interaction category for contactless HHI relation embedding in a multi-person scene. Specifically, as shown in Fig. 2, the proposed IRE-Net adopts not only the video image but also the involved multi-person spatial trajectories as input. We first use a short-term individual feature extraction method to obtain the feature for each subject, in which we use the GRU for temporal motion feature representation (Section 2). After that, we leverage a relation embedding strategy to integrate the individual features into a pairwise relation representation cube to model the mutual interactive relations among the multiple persons. This mainly models the spatial relations among the subjects in a short term. Then we apply a BiGRU network to aggregate the long-term feature in all short terms over the whole video. The obtained spatial-trajectory feature and long-term aggregation can better *model the temporal relation variations appearing in the contactless interactions* (Section 3). Finally, we fed the relation representation into a multi-head multi-task readout network to simultaneously identify the interactive subjects and recognize the interaction category (Section 4).

3.2 Individual feature extraction

Given a video sequence, we first split it into K segments with the same length. We sample N frames with short intervals, e.g., 5 frames, in each segment, and such short-interval sampled frames from each segment can better capture the whole process of rapid actions. In total, for a sequence of K segments, we sample $K \times N$ frames. We next present the feature extraction of each subject i from the sampled short-term N frames, i.e., the spatial-aware and appearance features of each subject.

Spatial-aware feature Unlike previous works mainly focusing on the human appearance or skeleton features, the proposed method models the multi-person *spatial distribution and variation* information by using the human location and trajectory. To capture the individual spatial-aware feature, we choose its 2D position coordinate and moving direction vector. For the ith subject in the tth frame, we denote its 2D position vector as $\mathbf{p}_i^t = (x_i, y_i) \in \mathcal{R}^2$. We further calculate the (short-term) moving direction vector $\mathbf{d}_i^t = \mathbf{p}_i^{t+l} - \mathbf{p}_i^{t-l} \in \mathcal{R}^2$ where we empirically set l as 4. To capture the temporal-

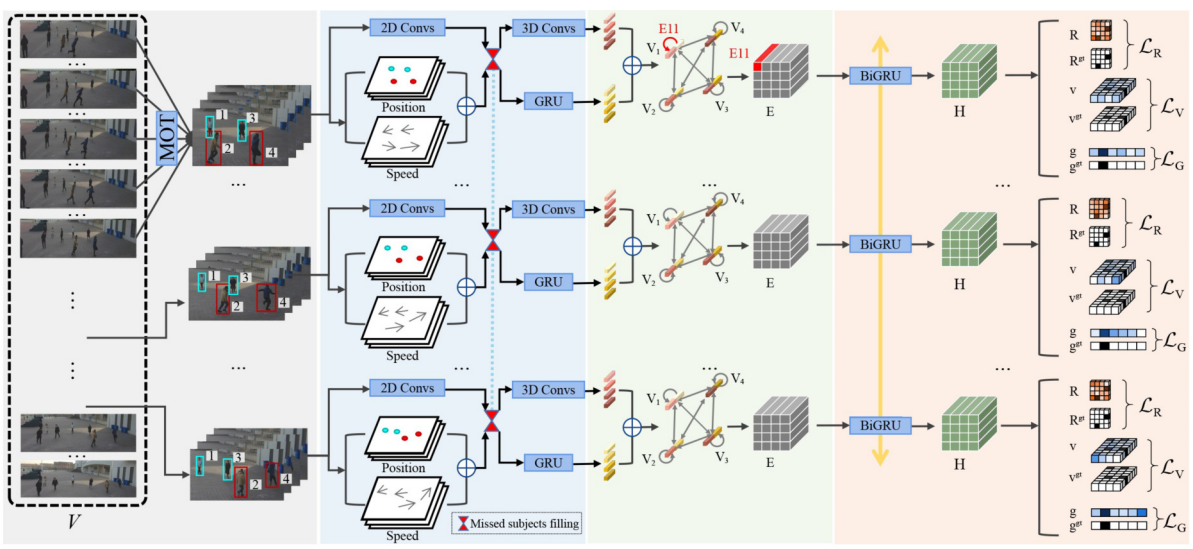


Fig. 2 Illustration of the proposed method for contactless interactive subject identification and interaction recognition

aware features, we further employ a single-layer GRU to integrate the positions of the same subject over the sampled N short-term frames in a segment as $\mathbf{h}_i^t = \text{GRU}(\text{concat}(\mathbf{p}_i^t, \mathbf{d}_i^t), \mathbf{h}_i^{t-1}), t = 1, 2, ..., N$, where $\mathbf{h}_i^t$ denotes the embedded temporal features on frame t, and $\mathbf{h}_i^1$ is initialized by the Xavier method [57]. This way, we use the final hidden state, i.e., $\mathbf{h}_i^N$ as the spatial-aware feature $\mathbf{F}_i^s$ of the subject i.

Appearance feature Following the feature extraction strategy used in [45], we employ Inception-v3 [58] to extract the multi-scale feature maps within each human bounding box. Then we use RoIAlign [59] to resize the extracted features into the same size. After that, we apply multiple 3D convolutional layers with a kernel size of $3 \times 1 \times 1$ to aggregate the temporal information over the N frames in a segment. After that, a fully connected (FC) layer is applied to get the appearance feature for each subject. For convenience, we denote the appearance feature of subject i in a segment as $\mathbf{F}_i^a$.

3.3 Interactive relation cube embedding

After extracting the individual spatial and appearance features, we aggregate all of them by a relation graph.

Node representation Given M subjects simultaneously appearing in the scene during a short-term segment. The graph node feature vector $\mathbf{F}_i$ of the subject, is constructed by concatenating spatial-aware and appearance features, i.e.,

$$\mathbf{F}_i = \text{concat}(\mathbf{F}_i^s, \mathbf{F}_i^a), \quad i \in 1, 2, \ldots, M. \tag{1}$$

Cube representation The graph edge feature vector from node i to j is set by $\mathbf{E}_{i,j} = \text{fc}(\text{concat}(\mathbf{F}_i, \mathbf{F}_j)), i, j \in 1, 2, \ldots, M$. The edge feature $\mathbf{E}_{i,j} \in \mathcal{R}^{N_e}$ is used to denote the co-embedding representation within each pair of subjects. As shown in Fig. 2, we propose a novel *interactive relation embedding cube* $\mathbf{E} \in \mathcal{R}^{M \times M \times N_e}$, that piles up all the $M \times M$ edge features, to model the pairwise relation among all the M subjects appearing in the scene during each short-term segment.

Long-term representation aggregation We then consider all the segments in the whole video, as shown in the right of Fig. 2. For a video with K segments and M subjects appearing in the scene, in each segment k, we can construct a short-term relation-aware feature cube denoting as $\mathbf{E}^k \in \mathcal{R}^{M \times M \times N_e}$ as discussed above. In total, we get K feature cubes $\mathbf{E}^k$, $k = 1, 2, \ldots, K$. We then apply a multi-layer bidirectional GRU to aggregate K short-term relation cubes into the long-term representation as

$$\mathbf{H}^k = \text{GRU}(\mathbf{E}^k \mathbf{H}^{k-1}) \in \mathcal{R}^{M \times M \times N_l}, k = 1, 2, \ldots, K. \tag{2}$$

As shown in Fig. 2, the aggregated feature cube $\mathbf{H}^k$ has the same structure as $\mathbf{E}^k$ but is with different number of channels, i.e., N_l. We denote $\mathbf{H}_{i,j}^k \in \mathcal{R}^{N_l}$ as the ith row, jth column vector of $\mathbf{H}^k$, which implicitly represents the relation between the ith subject and jth subject. In the next section, we elaborate on the use of $\mathbf{H}^k$ for interactor identification and interaction category recognition.

3.4 Multi-head multi-task interaction prediction

In this section, we denote $\mathbf{H}^k$ as $\mathbf{H}$ for simplicity, based on which we further discuss the output of interactor and interaction prediction via IRE-Net as shown in Fig. 3. The interaction category prediction is classified into two tasks – the individual level for each subject and the global level for whole video, respectively.

1) **Interactive relation predication** As shown in Fig. 3(a), we first apply a 2-layer FC operation to compress $\mathbf{H}$ along the channel dimension (Z-axis) and get $\text{fc}(\mathbf{H}) \in \mathcal{R}^{M \times M \times 1}$, then we flatten the obtained matrix into a one-dimension vector

$$\mathbf{v}_\mathbf{H} = \text{flatten}(\text{fc}(\mathbf{H}, \text{dim} = \text{Z})) \in \mathcal{R}^{M^2}, \tag{3}$$

on which we apply a softmax operation and then reshape the output vector into the matrix as the original order

$$\mathbf{R} = \text{reshape}(\text{softmax}(\mathbf{v}_\mathbf{H})). \tag{4}$$

The obtained $\mathbf{R} \in \mathcal{R}^{M \times M}$ can be taken as the *interactive relation probability matrix* among the M subjects, and we have $\sum_{p,q=1}^{M} \mathbf{R}(p, q) = 1$. Then we can define the interactive relation loss as

$$\mathcal{L}_R = \sum_t \text{L}_{bc}(\mathbf{R}_t, \mathbf{R}_t^{\text{gt}}), \tag{5}$$

where L_{bc} denotes the binary cross entropy loss, and we accumulate each frame t in a segment.

2) **Individual interaction category prediction** As shown in Fig. 3(b), to predict the interactive category (including all C interaction categories and the non-interaction) of each subject i in segment k, we first compress $\mathbf{H}$ along the second dimension (Y-axis) by taking the maximum value (the reason for using maximum here will be discussed later)

$$\mathbf{M} = \max(\mathbf{H}, \text{dim} = Y) \in \mathcal{R}^{M \times N_l}, \tag{6}$$

we then apply an FC layer $\text{fc} : M \times N_l \to M \times (C + 1)$ and the softmax operation, to define the *individual interaction category vector*

$$\mathbf{v}_i = \text{softmax}(\text{fc}(\mathbf{M})) \in \mathcal{R}^{C+1}, \tag{7}$$

which is to apply supervision of interaction category on each subject, including the category of non-interaction as

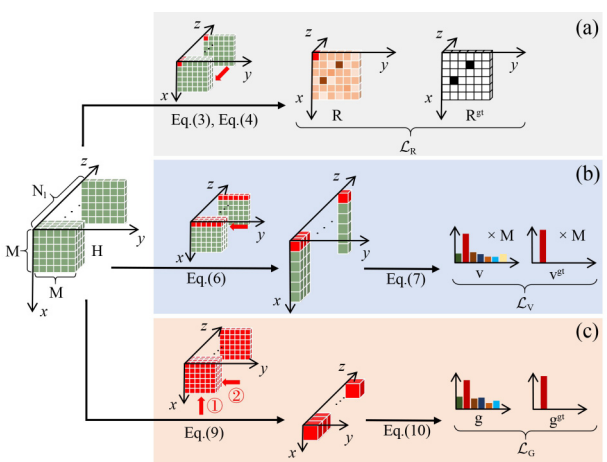


Fig. 3 Illustration of interaction prediction via IRE-Net. In the XOY plane, the three sub-tasks in this problem are modeled as a point, a line, and a face (with red color) of the proposed relation cube, respectively

$$\mathcal{L}_V = \sum_t \mathrm{L}_c(\mathbf{v}_t, \mathbf{v}_t^{\mathrm{gt}}), \tag{8}$$

where L_c denotes the cross entropy loss.

3) **Global interaction category prediction.** Finally, as shown in Fig. 3(c), we define a vector $\mathbf{g}^k$ to predict the interactive category of the whole video. Specifically, we compress $\mathbf{H}$ along both the first and second dimensions (*XOY*-plane) by take the maximum value and get

$$\mathbf{m} = \max(\mathbf{H}, \dim = XOY) \in \mathcal{R}^{N_l}, \tag{9}$$

then we apply an FC layer $\mathrm{fc} : N_l \to C$ followed by a softmax operation and get the *global interaction category vector* as

$$\mathbf{g} = \mathrm{softmax}(\mathrm{fc}(\mathbf{m})) \in \mathcal{R}^C, \tag{10}$$

where C denotes the number of interaction categories. The global interaction category prediction loss is defined as

$$\mathcal{L}_G = \mathrm{L}_c(\mathbf{g}, \mathbf{g}^{\mathrm{gt}}), \tag{11}$$

where L_c denotes the cross entropy loss.

Note that, existing works usually aggregate the features among the subjects in Eqs. (6) and (9) using the relation matrix $\mathbf{R}$ as weight, i.e., $\mathbf{M} = \mathrm{sum}(\mathbf{H} \odot \mathbf{R}, \dim = Y)$ and $\mathbf{m} = \mathrm{sum}(\mathbf{H} \odot \mathbf{R}, \dim = XOY)$. We use maximum operation for feature aggregation to make the features of the subjects with interaction more discriminative from others. It also alleviates that the predicted relation matrix with error may disturb the feature aggregation thus impact the interaction prediction. The ablation study in Section 3 will verify this point.

Discussion about the advantages of IRE-Net: 1) For the short-term individual feature extraction, we apply a dual-branch network to extract both the appearance and spatial features of each subject. Unlike previous human/group action recognition methods that mainly focus on the human appearance and skeleton features, we find that the spatial distribution and relation among the multiple subjects are important in our task. The proposed deep spatial-relation-aware features are helpful in finding the contactless HHI. 2) Since two subjects far from each other could have the contactless interaction, we propose a simple and effective strategy that uses a relation graph structure to integrate the individual features into a *pairwise-relation representation* for all the underlying interactors, which is important for the interactive relation discovery. 3) The *short-long-term feature aggregation* captures the short-term action-level (several frames) and long-term event-level (whole video) information, both of which are useful in our task.

IRE-Net novelly constructs the interactive relation cube, which is followed by a multi-task multi-head module to jointly handle: 1) interactive relation detection, 2) individual interaction recognition, 3) video-level interaction recognition. The representation of these three tasks could be simultaneously modeled from a point, a line, and a face of the same feature cube, as shown in Fig. 3. Thus, the total structure of the network stays simple and we can fully reuse parameters in handling these three tasks.

3.5 Implementation details

Loss function Taking the segment k for example, we employ four loss functions, i.e., $\mathcal{L}_R$, $\mathcal{L}_V$, $\mathcal{L}_G$, $\mathcal{L}_{\mathrm{tri}}$, and add them up as our overall loss function on segment k as $\mathcal{L}^k = \mathcal{L}_R + \mathcal{L}_V + \mathcal{L}_G + \mathcal{L}_{\mathrm{tri}}(\mathbf{m}_a, \mathbf{m}_p, \mathbf{m}_n)$, where $\mathcal{L}_R$, $\mathcal{L}_V$ and $\mathcal{L}_G$ are defined in Eqs. (5), (8) and (11), respectively and $\mathbf{m}_a$ denotes the compressed feature in Eq. (9) of the video a, and $\mathbf{m}_p, \mathbf{m}_n$ denote the features of the videos with the interaction categories that are the same as or different from the one in video a, respectively. $\mathcal{L}_{\mathrm{tri}}$ is the triplet loss function defined in [60]. The overall loss function on the whole video is the summation of $\mathcal{L}^k$ over all K segments.

Network training We adopt Inception-v3 [58] as backbone network, i.e., the 2D Convs in Fig. 2, to extract 1,024-dimensional features of each frame and then use RoIAlign [59] to extract and resize the feature of each bounding box to the size of 5×5 with 288-dimensional features. Due to the GPU memory limit, we train our model in two stages following [45]. First, we fine-tune the ImageNet pre-trained backbone network, on our datasets. We use a single frame randomly sampled from each video as input to train the backbone without other components. After training, we fix its parameters and used it to extract each frame's features in our dataset. Then we use the saved features and the bounding boxes to train the other parts of the proposed framework.

Network inference In the inference stage, we simultaneously obtain 1) interactive relation among the subjects, 2) the individual interaction category of each subject, and 3) the global interaction category predictions of the sequence. First, for the interactive relation, after getting the interactive relation matrix in Eq. (4), we calculate an interactive relation score $P_i \in [0, 1]$ to predict whether the ith subject is an interactor or not with $P_i = \sum_{j=1}^{M} \mathbf{R}_{i,j}$, which involves the interactive relation probability between ith subject and all other $M-1$ subjects. The interactive relation result is true if the predicted score P_i of subject i ranks in the top τ among all P_m for $m = 1, 2, \ldots, M$. In this paper, we consider the situation that there is at most one pair of HHI interactors in the scene thus we set $\tau = 2$. Second, the individual interaction category prediction of subject i can be obtained by $\mathrm{argmax}(\mathbf{v_i})$, where $\mathbf{v_i}$ is defined in Eq. (7). Third, the global interaction category prediction is calculated by $\mathrm{argmax}(\mathbf{g})$, where $\mathbf{g}$ is defined in Eq. (10). Note that, the interaction prediction obtained by Eq. (10) is computed for each segment, so we use a voting strategy to integrate the results from all segments and get the video-level prediction, under the assumption that each video contains one interaction label in our problem as clarified earlier. We do not apply the integration strategy for the interactive relation and individual interaction category because there usually exists subject ID shifts when associating humans in the MOT algorithm.

Experimental settings We use stochastic gradient descent with Adam optimizer to optimize the parameters. The batch size in training and testing is 16, and we train the framework for 100 epochs. We implement our method based on the PyTorch framework.

In this work, we obtain the person bounding boxes and over-time identity label by a state-of-the-art MOT algorithm FairMOT [61]. We also adopt a missed subject filling strategy for remedying the failed MOT results, which will be discussed in the following. We set the number of segments K as 7, and

the sampled frames in each segment N as 5, the dimensions of both appearance feature and spatial feature are 128. We set M as a fixed number of 30, which exceeds maximum number of subjects in each video in our dataset. We know that it is difficult to be satisfied that there are M subjects in the scene over the whole video, not to say all the videos in the dataset. This way, we propose a "Missed subject filling" strategy to handle the videos with different subjects for training together, which is presented in the following.

Missed subject filling strategy In the proposed method, we use a MOT algorithm for the over-time human association as the input, which may not always be accurate given the mutual occlusion. We find that the MOT errors may occur where one subject loses its original ID and obtains a new ID in the following frames. However, as discussed in Section 3, we assume there are M subjects in the scene over the whole videos. Given the failed subject tracking, or a subject may leaves or appears in the field of view of the camera, it is difficult to be satisfied that there are M subjects in the scene over the whole video, not to say all the videos, which contain different number of subjects in the dataset. This way, we proposed a simple missed subjects filling strategy to handle the above problem.

Our basic idea and rule are presented as below. 1) Exclusion rule: The different human IDs appearing in the same frame must belong to different persons. 2) Compatibility rule: The human IDs never simultaneously appearing in the same frame (for the whole video) may belong to the same person. Specifically, for example, the subject A is not detected from frame 1 to t_1, as shown in the top row in Fig. 4, thus we fill the absent range for such subject with another one from all the other possible subjects by comparing their feature similarity. By selecting smallest Manhattan distance between the features, we fill the subject B which appears in frame 1 to t_1 but does not appear in frame t_1+1 to T. Here the possible subjects filled for A are those never simultaneously appear together with subject A at the same time. For the subject without possible candidates to be filled, e.g., the subject C during t_2 to T in Fig. 4, we fill the absent range with its historical/future average feature. Note that, during the inference stage, we remove all the filled subjects to ensure that one subject is predicted with only one label.

4 Experiments

4.1 Dataset and metrics

Dataset collection We do not find publicly available human interaction datasets with contactless interaction in multi-person scenes for the proposed human interaction recognition and interactor localization. Therefore, we collect a new video dataset using the wearable camera GoPro in an outdoor scene. We define six daily contactless interactions including Coming Here (CO), Going Away (GO), Greeting (GR), Chasing (CH), Throw & Catch (TA), and Placing & Picking (PP).

- Coming Here (CO): Subject A swings the arm toward subject B for calling her/him. Subject B walks randomly until she/he notices A's signal and then walks toward A.
- Going Away (GO): Subject A swings the arm toward subject B for driving out her/him. Subject B walks randomly until she/he notices A's signal and then go away from A.
- Greeting (GR): Subject A swings the arm or uses specific gestures to greet subject B, who responds with a greeting back to A after she/he notices it.
- Chasing (CH): Subject A chases after subject B with a clearly faster speed than walking.
- Throw & Catch (TA): Subject A throws a small object to subject B, and subject B catches it by hand.
- Placing & Picking (PP): Subject A walks to a location and places a small object on the ground and then walks away. Then Subject B notices it and walks to the location, picks up the object, and then randomly walks away.

For data collection, we arrange 10 volunteers to perform the interactive activity for video collection – the persons not invovled in the interaction are doing sport or walking by. The data collection site is located at the entry of a playground in a campus, where students can walk by and play sports. In each video, the volunteers are randomly walking around, standing or talking with others without specific requirement. A director randomly asked two of the volunteers to perform one kind of the above-mentioned contactless interaction once. The two contactless interactors could be far from each other. The recorded videos are manually trimmed, and each trimmed sequence contains one category of contactless interaction, which was performed only once. This way, we collected 480 sequences in total. In different videos, the observers wearing GoPro cameras stand or sit at different places to watch the volunteers from different views. Note that, different from the movie shot, the observers are asked to cover most subjects rather than constantly focus on the interactors. The videos in the collected dataset have $1,920\times1,088$ spatial resolution in 30 fps.

For dataset annotations, we provide frame-level interactor identification labels and the interaction category label. Moreover, we provide the video-level interaction category labels, which is the category of the involved interactors. The bounding boxes are not annotated for the other people in the scene not joining any interaction, whose action labels are labeled as non-interaction.

Data statistics In total, we collect 480 videos with 103,634 frames in our dataset. Specifically, each category of interactions contains 80 sequences with different lengths. We

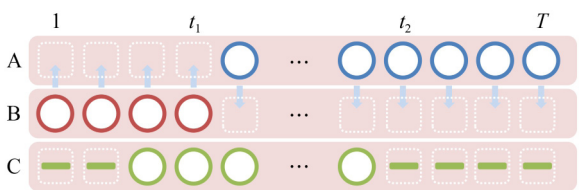


Fig. 4 Example cases of missed subjects filling in the proposed method. Here, a colored ball represents a feature vector of one subject at a frame. The empty block means the feature vector of one subject at this time is missing. The arrow means filling one feature to another place. The short bold line means the historical/future average feature

annotate each frame with bounding boxes of all the interactive subjects, and 206,841 human bounding boxes are annotated in total. The frame numbers for different interaction categories are imbalanced, which increases the challenge of our problem. We split the dataset into training and testing sets by 1 : 1, each of which contain 240 video sequences. Specifically, we arrange a crowd of people for video collection. Among them, we randomly select two subjects, e.g., Subject #A and #B, to do interaction. The two subject alternately do the interaction twice, take the "Coming Here" for example, we arrange #A walking toward #B and #B walking toward #A, respectively. This way, the generated two videos are used as the training data and testing data, respectively. Note that, the selection of subjects and interaction category is random. For the same subject pair with the same interaction, the action order is also different in the training and testing datasets. Although with identity overlap, our setting can guarantee that *the interaction category is not correlated to the human identity*. We select the same number of frames from every video sequence for all interaction categories. We show the statics of the proposed dataset in Table 1.

Comparison with previous dataset In this paper, we focus on a new problem, i.e., the contactless interactive activities in multi-person scenes, **which has not been studied in before human activity datasets**, like classical JHMDB, UCF101-24, AVA and recent EpicKitchen, ActEV, etc. Note the fact that most of the currrent HHI datasets focus on contact interactions or contactless interactions happened without many other neighbors to be potential interactive persons. Specifically, although some related public datasets, like AVA, JHMDB, UCF101-24, contain multiple-person scene and some interaction categories are similar to our defined contactless interactions, like watch to, talk to, but they still can not cover our interests. First, only a little part of their videos involve multi-person scene and HHI. More important, for the HHI in multi-person scene, due to the videos are mostly from movies and Internet, the spot view is totally different from the videos in surveillance video. In Table 2, we calculate the number of contactless-interaction categories and its ratio over all categories, and the average number of subjects per scene. We can see that *contactless interactions only count for a small percentage of action/interaction categories in prior datasets*. Also, the number of subjects in previous datasets is about two, which is much fewer than that in our dataset.

This paper studies the complex cases, which couldn't reply on the prior that interactors are close to each other of finding who are the interactors in the crowd. This is not well studied in the current HHI methods. Since the current datasets and open video source like Youtube videos and movies doesn't meet our requirement, we have to capture the videos by organizing volunteers. In our datasets, the distance between the two interactive subjects in our dataset are farther than two interacting actors in a movie shot. Besides, the spot view is from a GoPro wearable camera at an oblique downward viewing angle and the average number of persons in the scene is 7.2. Considering both the distance between interactors and the number of subjects in the scene, our task is more challenge in localizing the interactive subjects than those task in previous works.

Table 1 Statistics of the proposed dataset

Dataset	# Videos	# Frames	# Interactors
CO	80	15,064	30,128
GO	80	14,176	28,343
GR	80	8,754	17,508
CH	80	19,242	38,286
TA	80	8,914	17,800
PP	80	37,484	74,776
Training	240	52,500	104,887
Testing	240	51,134	101,954
Full	480	103,634	206,841

Table 2 Statistics and comparison of the proposed dataset and others

Dataset	# Type	Ratio	# Subjects
UT-interaction [6]	1	0.17	~2
ShakeFive2 [15]	0	0	2
SBU Kinect [7]	2	0.25	2
AVA [8]	2	0.25	1.8
Ours	6	1	9.6

For the data size, we clarify that current super-large-scale datasets mainly collect video clips from movies and video website, in which the photographer tend to focus on the key characters and the interactors are often conspicuous in the scene. This makes the interaction localization easy. To reflect the reality in real-world scenarios, e.g., video surveillance, we collect the videos by ourselves. Compared with other laboratory-collected video datasets, like SBU Kinect [7] (300 clips, 1-5 s), ShakeFive2 [15] (153 clips, 5-10 s), and CVID [1] (150 clips, 8-10 s), the scale of our dataset (480 clips, 3-8 s) is comparable.

Metrics We design the following metrics for evaluation.

- Metric I. Interactor identification. On each frame, we can use the relation matrix to represent the interactive relations between each pair of subject. For a predicted relation matrix $\hat{\mathbf{X}} \in \mathbb{R}^{M\times M}$ and a ground-truth adjacency matrix $\mathbf{X}^{\mathrm{gt}} \in \mathbb{R}^{M\times M}$, where M denotes the number of subjects, and $\mathbf{X}(i, j) = 1$ denotes the subjects i and j have interaction with each other, we define the interactor detection precision $\mathcal{P} = \frac{\sum \mathrm{AND}(\hat{\mathbf{X}},\mathbf{X}^{\mathrm{gt}})}{\sum \hat{\mathbf{X}}}$, and recall as $\mathcal{R} = \frac{\sum \mathrm{AND}(\hat{\mathbf{X}},\mathbf{X}^{\mathrm{gt}})}{\sum \mathbf{X}^{\mathrm{gt}}}$, where AND denotes the logical function. The numerator counts the true positive interactive relation among the subjects, while the denominators count the predicted and ground-truth interactive relations.
- Metric II. Video-level interaction recognition. We also evaluate the performance of interaction category prediction for the whole video. This can be regarded as a standard multi-label classification problem. Thus we utilize the classical $\mathcal{P}$, $\mathcal{R}$ and F_1 score $\mathcal{F}$ as metrics.
- Metric III. Individual-level interaction recognition. We evaluate the interaction category prediction performance for each subject with the interaction, which can also be regarded as a standard multi-label classification problem and we apply the metrics of $\mathcal{P}$, $\mathcal{R}$, and $\mathcal{F}$.

- Metric IV. Interactor identification & interaction recognition. We also use a comprehensive metric – Multiple Human Interaction Accuracy (MHIA) [1] for evaluation, which is defined as $\text{MHIA} = 1 - \frac{\sum_t \text{ms}_t + \sum_t \text{fp}_t + \sum_t \text{fc}_t}{\sum_t (\text{d}_t + \text{g}_t)}$, where ms_t and fp_t are the numbers of the false negative (missed) and false positive subjects for spatial-domain interactor detection at frame t, fc_t denotes the number of subjects with true interactor identification but false interaction category at frame t, and d_t/g_t represent the number of detected/ground-truth subjects with interaction at frame t. The overall MHIA score is calculated as the average value over all frames in a video.

4.2 Comparative results

Baselines As discussed in the related work, we can not find existing method that can *directly* handle the proposed problem. To compare our methods with others as much as possible, we consider various state-of-the-art approaches for human action/interaction recognition, social relation recognition (SDR), human-object interaction (HOI) detection, group activity recognition (GAR). Note that, to make the above methods applicable to our problem, we have to re-implement them with necessary modifications, which are presented in detail as below.

- Chance: A weak baseline that randomly predicts the interaction label for each subject and each video.
- X3D [62]: An efficient method for human action/interaction recognition, which takes a video containing the people performing an action as input and predicts the action label of the video. Since the official public code does not support the input of human bounding boxes, we feed the whole video to the X3D network for video-level interaction recognition.
- SlowFast [63]: A state-of-the-art human action/interaction recognition method. Following the setting of SlowFast in [63], we directly input the whole video to the SlowFast network without giving the bounding boxes of subjects. Therefore, just like X3D, it can only predict the video-level interaction recognition results in our task.
- SlowFast w box [63]: Following the setting of SlowFast with bounding box in [63], we input the video together with the human bounding boxes to the network. Note that, SlowFast can not directly output the interactive relations, but only the individual action of each subject. Therefore, we first use a voting strategy to estimate the video-level interaction category upon all the individual actions. Then we select the subjects with the top-two prediction of the individual interaction that is the same as the predicted global interaction category as the interactors.
- GR2N [48]: A state-of-the-art method for single image based social relation detection (SRD). We sample images from the video and input them into the network. The original network outputs predictions of the social relation between each pair of subjects, which are considered as the interaction relation and action of each subject. In order to obtain the prediction of the whole video, we add a GRU with the same setting as ours at the end of the network. We use the interaction relation as the weights to integrate the features of all subjects, which is used to predict the video-level interaction result.
- GPNN [50]: A recent method for human-object interaction (HOI) detection with graph neural network (GNN). The GPNN takes the whole video as input and outputs the interaction of each subject. We generate the interaction relation prediction where the subject with interaction as 1, and otherwise as 0. Then we obtain the video-level interaction predictions by the same way as GR2N.
- ARG [45], HiGCIN [64], Dynamic [65]: A series of state-of-the-art methods for group activity recognition (GAR) considering the human action/interaction. ARG is a classical method for GAR. It trains an Inception-v3 network to extract appearance features and uses graph neural networks (GNN) to predict each subject's (interactive) action and the group activity. We use the feature extraction backbone in ARG. To handle our task, we take the video's interaction label as the group activity label in ARG. We also use the supervision of each subject's action as in ARG. Besides, ARG contains a relation matrix in GNN, which is an $M \times M$ matrix that represents the interaction probability among M subjects in the scene. We use it to obtain the interaction relation prediction. The public code of ARG lacks the supervision of the relation matrix. We apply the supervision on the relation matrix as ours in Eq. (4) for a fair comparison.

Dynamic and HiGCIN use the spatial-temporal GNN and hierarchical GNN, respectively. Their original networks predict each subject's action and the group activity of the whole scene. We generate the interaction relation predictions where the subject with interactive action as 1, and otherwise as 0. Following the same training settings as we applied to ARG, we take the video's interaction label as the group activity label. Thus, we use the same three supervisions as the proposed method to train the networks in these methods.

As shown in Table 3, we evaluate our method with the comparison methods. First, in terms of the interaction detection task, we can see that a large part of the comparison methods show very poor performance, i.e., $\mathcal{F}$ lower than 30%. Among them, SlowFast, as an HHI recognition method, lacks the ability of modeling the relations among subjects in the multi-person scene. HOI detection method GPNN also generates poor results since it is designed to model the relation between human-object based on inherent priors and it is not appropriate for handling our problem that only involves humans. GAR method ARG though has the ability to model the interactive relations among the subjects, it is still much lower than the proposed method. Although ARG provides a high recall score, indicating that it predicts a wide range of subjects as interactive ones, its precision is very poor. This

Table 3 Comparative results of interactor identification, video interaction recognition and subject interaction recognition (%)

Method	Interactor Ind.			Vid. Interaction Rec.			Sub. Interaction Rec.			Overall
	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}$	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}$	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}$	MHIA
Chance	15.1	6.3	8.8	15.6	15.8	15.6	2.3	13.6	4.0	4.9
X3D [62]	–	–	–	60.4	22.4	32.5	–	–	–	–
SlowFast [63]	–	–	–	59.3	21.1	30.9	–	–	–	–
SlowFast w box [63]	12.5	13.2	12.2	45.1	44.2	41.6	12.4	16.3	14.1	12.0
ARG [45]	14.6	**75.3**	24.3	58.7	58.8	58.2	8.8	15.2	11.2	18.2
GR2N [48]	53.9	53.3	53.6	51.3	52.9	50.3	17.9	55.3	27.1	40.0
GPNN [50]	20.0	20.6	20.3	55.1	52.5	48.2	9.1	**57.5**	15.7	15.7
HiGCIN [64]	50.4	51.9	51.1	63.1	61.2	60.8	21.9	54.6	31.2	41.4
Dynamic [65]	40.3	41.5	40.9	55.5	55.8	53.1	19.6	27.4	22.8	30.4
Ours	**65.2**	48.1	**55.3**	**63.8**	**65.0**	**64.2**	**40.0**	42.0	**41.0**	**44.2**

may be beneficial for group activity recognition but not effective in our task. For the later coming GAR methods Dynamic and HiGCIN, both of them have paid attention to precisely modeling the relations among subjects by spatial-temporal context, and achieved considerably good results. We can also see that the social relation recognition method GR2N, in part, has the potential to model the interactive relation. However, the above-mentioned results are still worse than ours.

For the video-level interaction category recognition, many baseline approaches show comparative performance. Specifically, SlowFast with bounding boxes is much better than without bounding boxes in recognizing video-level interactions, but it still performs worse than the GAR approaches and the proposed method. This implies that human action recognition methods that only consider individual information can not model the relation among multiple subjects, and therefore, they are unsuitable for our task. GR2N and GPNN considering the mutual relations among the subjects produce the acceptable performance on this sub-problem. Some GAR approaches including ARG, HiGCIN and Dynamic all show good video-level interaction recognition performance since this task is similar with the original GAR task, which, however, is still lower than the proposed method in F_1 score.

For the individual-level interaction recognition, which can be regarded as the integrated evaluation of interactor detection and interaction recognition, we can see that all of the comparison methods perform poorly and have a large margin to the proposed method. This demonstrates that the comparison methods can not well handle the proposed problem directly, which is different from the previous tasks. From the last column in Table 3, we can see that our method also outperforms all the comparison methods using the comprehensive metric MHIA.

4.3 Ablation study

To evaluate the effectiveness of our essential model components, we derive the following variants of our method:

- w/o Spatial/Appearance: Removing the spatial-aware features $\mathbf{F}_i^s$ / appearance features $\mathbf{F}_i^a$ in Eq. (1).
- w/o Short-term Sap.: Following frame sampling strategy in TSN [66], we set the short-term sampling frames in each segment in the way that N=1.
- w/o Long-term Agr.: Removing the GRU for long-term feature aggregation, namely we replace $\mathbf{H}_k$ with $\mathbf{E}_k$ in Eq. (2).
- w/o Cube: We replace the proposed interactive relation cube structure with the classical graph convolutional network (GCN) commonly used in previous relation modeling works, like ARG [45].
- w/o Filling: Removing the missing subjects filling decribed in Section 5. All the features of missed subjects are filled with zeros.
- w/o triplet: Removing the triplet loss in training.
- w $\mathbf{R}$ weight: We use the interactive relation matrix $\mathbf{R}$ as the weight map for interaction prediction, i.e., we change Eqs. (6) and (9) into $\mathbf{M} = \text{sum}(\mathbf{H} \odot \mathbf{R}, \text{dim} = Y)$ and $\mathbf{m} = \text{sum}(\mathbf{H} \odot \mathbf{R}, \text{dim} = XOY)$.

As shown at the top of Table 4, we first remove the spatial-aware features and appearance features, respectively, to study their impact to the final performance. After removing them, both performances decrease in most of the evaluating metrics. We can see that, without spatial-aware features, the performance of video-level interaction recognition of the proposed method drops 4.1% in F_1 score. All the F_1 scores for interactor detection and individual-level interaction recognition, and the MHIA score also decrease to some extent. This demonstrates that the spatial-aware feature is beneficial for distinguishing the contactless interaction in the multi-person scene. Surprisingly, without the appearance feature, although the performance decreases in most evaluation metrics, the proposed method could also achieve a considerable high accuracy on recognizing the video-level interaction category, even higher than ARG without interactive relation supervision.

We also evaluate the performance of the proposed method without long-term temporal information aggregation, and short-term multi-frame sampling. Without short-term sampling, the performance decreases by a small margin. Without long-term temporal modeling, the performance decreases by a large margin – the F_1 scores in the three tasks drop by 7.1%, 1.4%, 7.5%, respectively, and the MHIA also drops by 5.0%. This demonstrates that the long-term temporal aggregation is essential for our tasks, since the contactless interactive activities usually take a duration of time. Also, without the relation cube, the overall performance especially of the interactor identification becomes very poor, which verifies the effectiveness of the proposed cube for human

Table 4 Ablation study results of the proposed method for the three tasks (%)

Method	Interactor Ind.			Vid. Interaction Rec.			Sub. Interaction Rec.			Overall
	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}$	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}$	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}$	MHIA
w/o Spatial	64.1	46.3	53.7	60.4	62.5	60.1	35.4	45.6	39.8	42.8
w/o Appearance	41.5	41.9	41.7	58.1	57.9	57.6	15.7	58.6	24.7	38.2
w/o Long-term Agr.	54.9	43.0	48.2	63.2	63.3	62.8	24.7	52.0	33.5	39.2
w/o Short-term Sap.	62.3	47.5	53.4	64.7	64.6	63.7	38.6	41.2	39.8	43.7
w/o Cube	21.1	21.7	21.4	65.0	65.4	65.0	38.0	39.8	38.9	16.2
w/o Filling	66.2	46.2	54.4	64.1	64.6	63.5	26.7	54.3	35.8	42.7
w/o triplet	66.1	50.4	57.2	60.2	61.3	60.3	49.8	28.6	36.3	43.9
w $\mathbf{R}$ weight	61.6	45.2	52.1	57.7	58.3	55.9	36.2	42.4	39.0	43.3
Ours	65.2	48.1	**55.3**	63.8	65.0	**64.2**	40.0	42.0	**41.0**	**44.2**

relation representation.

For other components in our framework, we find that, without missed subject filling, the performance shows a slight decrease in all tasks. The situation is similar when removing the triplet loss. In the ablation study of using the interactive relation matrix as the weight map for predicting the interaction, i.e., w $\mathbf{R}$ weight, we can see a noticeable performance drop in the video-level interaction recognition task. This can be explained that the error in interactive relation detection may influence the interaction recognition.

5 Discussion

Generalization and extension. This work takes the first step to extend the current HHI research to HHI in more practical multi-person scene, which could benefit the real-world video surveillance and multimedia analysis. We aim to bridge the HHI problem and multi-person activity/relation analysis problem and lead to more comprehensive multi-person human activity understanding. Specifically, two-person HHI, as the most common human interactions, is important and has been widely studied in the community. In this paper, we extend the two-person HHI in two directions, i.e., from contacted ones to contactless, and from two-person scenes to multi-person scenes. This is more practical in the real world.

This work focuses on the scenes involving the contactless HHI, not be specifically studied before, for interaction understanding in the multi-person scenes. Based on it, for more complex human interaction tasks in crowded scenes, e.g., multiple pairs of various HHIs or multi-human interactive activities, they could be regarded as the combination or extension of the proposed problem.

Moreover, the proposed framework could be easily extended to more complex human interaction tasks, e.g., multiple (two-person) HHIs or the multi-person subgroup interaction recognition, by changing the aggregation operations upon the proposed interactive relation cube. Specifically, as shown in Fig. 3, we predict the interactive relation using each point in the cube as in Fig. 3 (a). For more complex human interactions, we can replace the "softmax" in Eq. (4) with a sigmoid function, which can generate the interactive relation of multiple pairwise HHIs or the multi-person interaction. For the individual interaction category prediction shown in Fig. 3 (b), we predict the interaction of all individuals, which can be directly used for the above scenes. Similarly, for the global interaction category prediction in Fig. 3 (c), in the more complex scenes, the prediction becomes a multi-category classification problem. We can also replace the "softmax" in Eq. (10) with a sigmoid function to get the multi-category prediction results. In conclusion, the proposed method address a fundamental problem of the HHI in multi-person scene, which is also easy to extended to more complex scene with appropriate modifications.

Limitation We do not conduct the experiments on the public datasets since they can not be used for our task and this work does not focus on the competition on the traditional HHI without contactless interactions or involving the subjects without interaction. Through the effort, we want to make a little attempt to provide the community the first dataset to explicitly address the proposed new but helpful task.

One limitation of this work maybe that the dataset is relatively small compared with existing large video datasets. Note that, the current large-scale datasets are mainly collected from movies and video website, where the interaction localization are easy and they are not applicable for our problem. To reflect the reality in real-world scenarios, like the video surveillance, we collect the videos by ourselves. Based on this *first moderately large dataset*, we are willing to further expand it by including more collected videos and interaction categories, including the mix of contact and contactless HHIs, in the future.

6 Conclusion

In this paper, we have studied a new problem of detecting contactless human interactions from multi-person scenes. Specifically, we proposed a new IRE-Net that combines the appearance and spatial features for representation and uses an individual-to-group short-to-long aggregation for interaction-aware pairwise-relation embedding, followed by a multi-head multi-task module to jointly identify the interactive subjects and recognize the interaction category. We also collected a new dataset for evaluating the proposed method. Through the above efforts, we hope to provide the resources, including the dataset and baselines, for studying this new problem, which extends the HHI recognition, from contacted ones to contactless, and from two-person scenes to multi-person scenes. In the future, we also plan to make use of the multi-camera collaboration for more effective multi-human activity analysis [1,67,68].

Acknowledgements This work was supported by the National Natural Science Foundation of China (NSFC) (Grant Nos. 62072334, U1803264).

Competing interests The authors declare that they have no competing interests or financial conflicts to disclose.

References

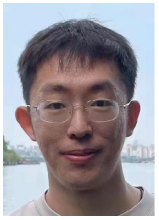
1. Zhao J, Han R, Gan Y, Wan L, Feng W, Wang S. Human identification and interaction detection in cross-view multi-person videos with wearable cameras. In: Proceedings of the 28th ACM International Conference on Multimedia. 2020
2. Li G, Qu W, Huang Q. A multiple targets appearance tracker based on object interaction models. IEEE Transactions on Circuits and Systems for Video Technology, 2012, 22(3): 450–464
3. Liang J, Jiang L, Niebles J C, Hauptmann A G, Li F F. Peeking into the future: predicting future person activities and locations in videos. In: Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2019
4. Mehran R, Oyama A, Shah M. Abnormal crowd behavior detection using social force model. In: Proceedings of 2009 IEEE Conference on Computer Vision and Pattern Recognition. 2009
5. Han R, Zhao J, Feng W, Gan Y, Wan L, Wang S. Complementary-view co-interest person detection. In: Proceedings of the 28th ACM International Conference on Multimedia. 2020
6. Ryoo M S, Aggarwal J K. Interaction dataset, ICPR 2010 contest on semantic description of human activities (SDHA 2010). See cvrc.ece.utexas.edu/SDHA2010/Human_Interaction.html website, 2010
7. Yun K, Honorio J, Chattopadhyay D, Berg T L, Samaras D. Two-person interaction detection using body-pose features and multiple instance learning. In: Proceedings of 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. 2012
8. Gu C, Sun C, Ross D A, Vondrick C, Pantofaru C, Li Y, Vijayanarasimhan S, Toderici G, Ricco S, Sukthankar R, Schmid C, Malik J. AVA: a video dataset of spatio-temporally localized atomic visual actions. In: Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018
9. Han R, Feng W, Zhang Y, Zhao J, Wang S. Multiple human association and tracking from egocentric and complementary top views. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(9): 5225–5242
10. Han R, Zhang Y, Feng W, Gong C, Zhang X, Zhao J, Wan L, Wang S. Multiple human association between top and horizontal views by matching subjects' spatial distributions. 2019, arXiv preprint arXiv: 1907.11458
11. Han R, Feng W, Zhao J, Niu Z, Zhang Y, Wan L, Wang S. Complementary-view multiple human tracking. In: Proceedings of the 34th AAAI Conference on Artificial Intelligence. 2020
12. Carreira J, Noland E, Hillier C, Zisserman A. A short note on the kinetics-700 human action dataset. 2019, arXiv preprint arXiv: 1907.06987
13. Kay W, Carreira J, Simonyan K, Zhang B, Hillier C, Vijayanarasimhan S, Viola F, Green T, Back T, Natsev P, Suleyman M, Zisserman A. The kinetics human action video dataset. 2017, arXiv preprint arXiv: 1907.06987
14. Kong Y, Jia Y, Fu Y. Learning human interaction by interactive phrases. In: Proceedings of the 12th European Conference on Computer Vision. 2012
15. Van Gemeren C, Poppe R, Veltkamp R C. Spatio-temporal detection of fine-grained dyadic human interactions. In: Proceedings of the 7th International Workshop on Human Behavior Understanding. 2016
16. Taylor G W, Fergus R, LeCun Y, Bregler C. Convolutional learning of spatio-temporal features. In: Proceedings of the 11th European Conference on Computer Vision. 2010
17. Tran D, Bourdev L, Fergus R, Torresani L, Paluri M. Learning spatiotemporal features with 3d convolutional networks. In: Proceedings of 2015 IEEE International Conference on Computer Vision (ICCV). 2015
18. Carreira J, Zisserman A. Quo vadis, action recognition? A new model and the kinetics dataset. In: Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017
19. Zhang C, Zou Y, Chen G, Gan L. PAN: persistent appearance network with an efficient motion cue for fast action recognition. In: Proceedings of the 27th ACM International Conference on Multimedia. 2019
20. Wang Z, Liu S, Zhang J, Chen S, Guan Q. A spatio-temporal crf for human interaction understanding. IEEE Transactions on Circuits and Systems for Video Technology, 2017, 27(8): 1647–1660
21. Motiian S, Siyahjani F, Almohsen R, Doretto G. Online human interaction detection and recognition with multiple cameras. IEEE Transactions on Circuits and Systems for Video Technology, 2017, 27(3): 649–663
22. Song S, Lan C, Xing J, Zeng W, Liu J. An end-to-end spatio-temporal attention model for human action recognition from skeleton data. In: Proceedings of the 31st AAAI Conference on Artificial Intelligence. 2017
23. Gao X, Hu W, Tang J, Liu J, Guo Z. Optimized skeleton-based action recognition via sparsified graph regression. In: Proceedings of the 27th ACM International Conference on Multimedia. 2019
24. Tang Y, Tian Y, Lu J, Li P, Zhou J. Deep progressive reinforcement learning for skeleton-based action recognition. In: Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018
25. Wang Z, Ge J, Guo D, Zhang J, Lei Y, Chen S. Human interaction understanding with joint graph decomposition and node labeling. IEEE Transactions on Image Processing, 2021, 30: 6240–6254
26. Feichtenhofer C, Pinz A, Wildes R P. Spatiotemporal residual networks for video action recognition. In: Proceedings of the 30th International Conference on Neural Information Processing Systems. 2016
27. Tran D, Wang H, Torresani L, Ray J, LeCun Y, Paluri M. A closer look at spatiotemporal convolutions for action recognition. In: Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018
28. Qiu Z, Yao T, Mei T. Learning spatio-temporal representation with pseudo-3D residual networks. In: Proceedings of 2017 IEEE International Conference on Computer Vision (ICCV). 2017
29. Wang H, Schmid C. Action recognition with improved trajectories. In: Proceedings of 2013 IEEE International Conference on Computer Vision. 2013
30. Wang L, Qiao Y, Tang X. Action recognition with trajectory-pooled deep-convolutional descriptors. In: Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2015
31. Lee D G, Lee S W. Human interaction recognition framework based on interacting body part attention. Pattern Recognition, 2022, 128: 108645
32. Tu H, Xu R, Chi R, Peng Y. Multiperson interactive activity recognition based on interaction relation model. Journal of Mathematics, 2021, 2021: 5576369
33. Verma A, Meenpal T, Acharya B. Multiperson interaction recognition in images: a body keypoint based feature image analysis. Computational Intelligence, 2021, 37(1): 461–483
34. Patron-Perez A, Marszalek M, Reid I, Zisserman A. Structured learning of human interactions in TV shows. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 34(12): 2441–2453
35. Zhao H, Torralba A, Torresani L, Yan Z. HACS: human action clips and segments dataset for recognition and temporal localization. In: Proceedings of 2019 IEEE/CVF International Conference on Computer

Vision (ICCV). 2019
36. Joo H, Liu H, Tan L, Gui L, Nabbe B, Matthews I, Kanade T, Nobuhara S, Sheikh Y. Panoptic studio: a massively multiview system for social motion capture. In: Proceedings of 2015 IEEE International Conference on Computer Vision (ICCV). 2015
37. Ehsanpour M, Saleh F, Savarese S, Reid I, Rezatofighi H. JRDB-Act: a large-scale dataset for spatio-temporal action, social group and activity detection. In: Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2022
38. Li J, Han R, Yan H, Qian Z, Feng W, Wang S. Self-supervised social relation representation for human group detection. In: Proceedings of the 17th European Conference on Computer Vision. 2022
39. Han R, Yan H, Li J, Wang S, Feng W, Wang S. Panoramic human activity recognition. In: Proceedings of the 17th European Conference on Computer Vision. 2022
40. Shu T, Todorovic S, Zhu S C. CERN: confidence-energy recurrent network for group activity recognition. In: Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017
41. Shu X, Tang J, Qi G, Liu W, Yang J. Hierarchical long short-term concurrent memory for human interaction recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 43(3): 1110–1118
42. Zhang P, Tang Y, Hu J F, Zheng W S. Fast collective activity recognition under weak supervision. IEEE Transactions on Image Processing, 2020, 29: 29–43
43. Yuan H, Ni D. Learning visual context for group activity recognition. In: Proceedings of the 35th AAAI Conference on Artificial Intelligence. 2021
44. Yan R, Tang J, Shu X, Li Z, Tian Q. Participation-contributed temporal dynamic model for group activity recognition. In: Proceedings of the 26th ACM International Conference on Multimedia. 2018
45. Wu J, Wang L, Wang L, Guo J, Wu G. Learning actor relation graphs for group activity recognition. In: Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2019
46. Choi W, Shahid K, Savarese S. What are they doing?: Collective activity classification using spatio-temporal relationship among people. In: Proceedings of the 12th IEEE International Conference on Computer Vision Workshops, ICCV Workshops. 2009
47. Ibrahim M S, Muralidharan S, Deng Z, Vahdat A, Mori G. A hierarchical deep temporal model for group activity recognition. In: Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016
48. Li W, Duan Y, Lu J, Feng J, Zhou J. Graph-based social relation reasoning. In: Proceedings of the 16th European Conference on Computer Vision. 2020
49. Li J, Wong Y, Zhao Q, Kankanhalli M S. Visual social relationship recognition. International Journal of Computer Vision, 2020, 128(6): 1750–1764
50. Qi S, Wang W, Jia B, Shen J, Zhu S C. Learning human-object interactions by graph parsing neural networks. In: Proceedings of the 15th European Conference on Computer Vision. 2018
51. Zhong X, Ding C, Qu X, Tao D. Polysemy deciphering network for robust human–object interaction detection. International Journal of Computer Vision, 2021, 129(6): 1910–1929
52. Qiao T, Men Q, Li F W, Kubotani Y, Morishima S, Shum H P H. Geometric features informed multi-person human-object interaction recognition in videos. In: Proceedings of the 17th European Conference on Computer Vision. 2022
53. Bai L, Chen F, Tian Y. Automatically detecting human-object interaction by an instance part-level attention deep framework. Pattern Recognition, 2023, 134: 109110
54. Li F, Wang S, Wang S, Zhang L. Human-object interaction detection: a survey of deep learning-based methods. In: Proceedings of the 2nd CAAI International Conference on Artificial Intelligence. 2022
55. Antoun M, Asmar D. Human object interaction detection: design and survey. Image and Vision Computing, 2023, 130: 104617
56. Lim J, Baskaran V M, Lim J M Y, Wong K, See J, Tistarelli M. ERNet: an efficient and reliable human-object interaction detection network. IEEE Transactions on Image Processing, 2023, 32: 964–979
57. Glorot X, Bengio Y. Understanding the difficulty of training deep feedforward neural networks. In: Proceedings of the 13th International Conference on Artificial Intelligence and Statistics. 2010
58. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. In: Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016
59. He K M, Gkioxari G, Dollár P, Girshick R. Mask R-CNN. In: Proceedings of 2017 IEEE International Conference on Computer Vision (ICCV). 2017
60. Schroff F, Kalenichenko D, Philbin J. FaceNet: a unified embedding for face recognition and clustering. In: Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2015
61. Zhang Y, Wang C, Wang X, Zeng W, Liu W. FairMOT: on the fairness of detection and re-identification in multiple object tracking. International Journal of Computer Vision, 2021, 129(11): 3069–3087
62. Feichtenhofer C. X3D: expanding architectures for efficient video recognition. In: Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2020
63. Feichtenhofer C, Fan H, Malik J, He K. SlowFast networks for video recognition. In: Proceedings of 2019 IEEE/CVF International Conference on Computer Vision (ICCV). 2019
64. Yan R, Xie L, Tang J, Shu X, Tian Q. HiGCIN: hierarchical graph-based cross inference network for group activity recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(6): 6955–6968
65. Yuan H, Ni D, Wang M. Spatio-temporal dynamic inference network for group activity recognition. In: Proceedings of 2021 IEEE/CVF International Conference on Computer Vision (ICCV). 2021
66. Wang L, Xiong Y, Wang Z, Qiao Y, Lin D, Tang X, Van Gool L. Temporal segment networks: towards good practices for deep action recognition. In: Proceedings of the 14th European Conference on Computer Vision. 2016
67. Han R, Gan Y, Li J, Wang F, Feng W, Wang S. Connecting the complementary-view videos: joint camera identification and subject association. In: Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022
68. Han R, Gan Y, Wang L, Li N, Feng W, Wang S. Relating view directions of complementary-view mobile cameras via the human shadow. International Journal of Computer Vision, 2023, 131(5): 1106–1121

Jiacheng Li received the BS degree in computer sciense and technology from Beijing University of Chemical Technology, China in 2019, and the ME degree in computer sciense and technology from Tianjin University, China in 2022. His major research interest is visual intelligence, specifically including multi-object interaction and social relation discovery.

Ruize Han received the BS degree in mathematics and applied mathematics from Hebei University of Technology, China in 2016, the ME and PhD degrees in computer sciense and technology from Tianjin University, China in 2019 and 2023, respectively. His major research interest is visual intelligence, specifically including multi-camera video collaborative analysis and multi-human activity understanding. He was also interested in solving preventive conservation problems of cultural heritages via artificial intelligence.

Wei Feng received the PhD degree in computer science from City University of Hong Kong, China in 2008. From 2008 to 2010, he was a research fellow at the Chinese University of Hong Kong, China and City University of Hong Kong, China. He is now a Professor at the School of Computer Science and Technology, College of Computing and Intelligence, Tianjin University, China. His major research interests are active robotic vision and visual intelligence. Recently, he focuses on solving preventive conservation problems of cultural heritages via computer vision and machine learning. He is the Associate Editor of Neurocomputing and Journal of Ambient Intelligence and Humanized Computing.



Haomin Yan received the BE degree in the School of Electrical and Information Engineering and the ME degree in computer technology from Tianjin University, China in 2020 and 2023, respectively. His research interests focus on multi-human action analysis, specially for the weakly supervised individual action detection and social group activity detection.

Song Wang received the PhD degree in electrical and computer engineering from the University of Illinois at Urbana Champaign (UIUC), USA in 2002. He was a Research Assistant with the Image Formation and Processing Group, Beckman Institute, UIUC, USA from 1998 to 2002. In 2002, he joined the Department of Computer Science and Engineering, University of South Carolina, USA, where he is currently a Professor. His current research interests include computer vision, image processing, and machine learning. Dr. Wang is currently serving as the Publicity/Web Portal Chair of the Technical Committee of Pattern Analysis and Machine Intelligence of the IEEE Computer Society, an Associate Editor of IEEE Transaction on Pattern Analysis and Machine Intelligence, IEEE Transaction on Multimedia and Pattern Recognition Letters. He is a Senior Member of the IEEE and a member of the IEEE Computer Society.