Human Identification and Interaction Detection in Cross-View Multi-Person Videos with Wearable Cameras

Jiewen Zhao*
Ruize Han*
Yiyang Gan*
College of Intelligence and Computing, Tianjin University
{zhaojw,han_ruize,realgump}@tju.edu.cn

Liang Wan†
College of Intelligence and Computing,
Tianjin University
lwan@tju.edu.cn

Wei Feng
College of Intelligence and Computing,
Tianjin University
wfeng@tju.edu.cn

Song Wang†
Tianjin University
University of South Carolina
songwang@cec.sc.edu

ABSTRACT

Compared to a single fixed camera, multiple moving cameras, e.g., those worn by people, can better capture the human interactive and group activities in a scene, by providing multiple, flexible and possibly complementary views of the involved people. In this setting the actual promotion of activity detection is highly dependent on the effective correlation and collaborative analysis of multiple videos taken by different wearable cameras, which is highly challenging given the time-varying view differences across different cameras and mutual occlusion of people in each video. By focusing on two wearable cameras and the interactive activities that involve only two people, in this paper we develop a new approach that can simultaneously: (i) identify the same persons across the two videos, (ii) detect the interactive activities of interest, including their occurrence intervals and involved people, and (iii) recognize the category of each interactive activity. Specifically, we represent each video by a graph, with detected persons as nodes, and propose a unified Graph Neural Network (GNN) based framework to jointly solve the above three problems. A graph matching network is developed for identifying the same persons across the two videos and a graph inference network is then used for detecting the human interactions. We also build a new video dataset, which provides a benchmark for this study, and conduct extensive experiments to validate the effectiveness and superiority of the proposed method.

CCS CONCEPTS

• **Computing methodologies** → **Activity recognition and understanding**.

KEYWORDS

human identification; interaction detection; multi-view video analysis; wearable camera

ACM Reference Format:
Jiewen Zhao, Ruize Han, Yiyang Gan, Liang Wan, Wei Feng, and Song Wang. 2020. Human Identification and Interaction Detection in Cross-View Multi-Person Videos with Wearable Cameras. In *Proceedings of the 28th ACM International Conference on Multimedia (MM '20), October 12–16, 2020, Seattle, WA, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3394171.3413903

1 INTRODUCTION

In recent years, video-based interactive and group activity analysis has attracted the interest of many researchers in the computer vision and multimedia communities. Most existing methods are developed for the videos captured by a single fixed camera, which can only cover specified and limited regions. In many scenarios with many people around, e.g., in an outdoor gathering, it can be difficult to identify the activities of interest over time from a single video taken by a fixed camera with pre-specified view – some involved people, referred to as subjects in this paper, may be out of the camera view or occluded by others from time to time. This limitation can be well addressed by using multiple moving cameras, such as cameras worn by several people in or near the scene [10–12]. For example, in an outdoor gathering, security personnels can wear cameras such as Google Glass or GoPro and walk around to record what they see over time. The resulting multiple wearable-camera videos can provide much more information than a fixed-camera video for interactive and group activity analysis.

To better observe a group of people on the ground, the location and perspective of cameras are very important. Figure 1(a) and 1(b) show the same scene simultaneously captured by two cameras from different locations and views. The same scene captured from View-1 (top) and View-2 (bottom) is totally different. In the red bounding box in Figure 1(a), we can not identify the ongoing activity because

*Equal contribution and co-first authors.
†Co-corresponding authors.
Authors are also with the Key Research Center for Surface Monitoring and Analysis of Cultural Relics, SACH, China.

MM '20, October 12–16, 2020, Seattle, WA, USA

ACM ISBN 978-1-4503-7988-5/20/10…$15.00
https://doi.org/10.1145/3394171.3413903

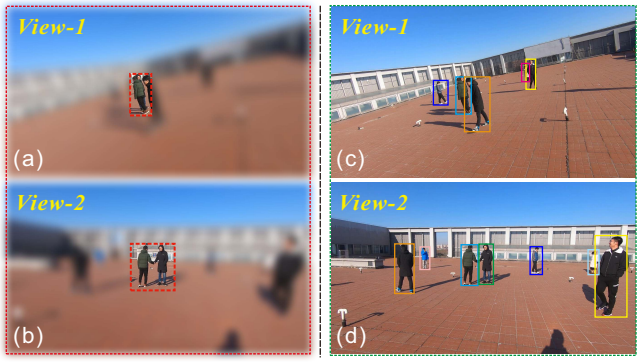

Figure 1: An illustration of our problem. Left: Highlighted region focused on the interactive subjects (a-b). Right: Whole scene with more subjects besides the interactive ones. The same subjects in different views are indicated with identical-color boxes (c-d).

of the mutual occlusion. From another view as shown in Figure 1(b), we can clearly observe the interactive activity of hand shaking between two subjects. This indicates that the multiple videos from different-view cameras may provide complementary information for better activity analysis. The use of multiple wearable cameras not only generates videos from different views, but also from time-varying views which can better capture the activities of interest. In this paper, we propose to correlate these videos and conduct combined video analysis for human interaction detection, by considering the complexities caused by cross-video view difference and within-video mutual occlusions, as shown in Figure 1(c-d).

More specifically, we propose to address the following three important tasks simultaneously: **Cross-view person identification (Task I)** – identifying the same persons across multiple videos captured by different wearable cameras. **Human interaction detection (Task II)** – detecting the human interactions of interest, including their occurrence intervals and involved subjects. **Human interaction recognition (Task III)** – recognizing the category of each detected interactive activity. The mutual dependence and promotion of these three tasks motivate the design of a unified framework to address them simultaneously. On one hand, accurate cross-view person identification (Task 1) is essential for combined video analysis for better detecting and recognizing the human interactions from multiple videos (Tasks II & III). On the other hand, the correct detection and recognition of human interactions (Tasks II & III) can also benefit the cross-view person identification (Task I). Moreover, the detection and recognition of human activities (Tasks II & III) in a video are well known to be two highly correlated tasks.

For simplicity, in this paper, we focus on two wearable-camera videos and the interactive activities that only involve two subjects, e.g., hand shaking. Compared to many existing works, our problem formulation is different in two perspectives. First, most researches on human-human interactions (HHI) [32, 35, 37, 46] were predominantly focused on recognition instead of detection. In these works, HHI category labels are usually assigned to a whole video without specifying the occurrence intervals, i.e., the starting and ending time of each interaction. In this paper, we not only recognize the HHI category, but also detect the starting and ending time of each activity. Second, in most of the prior works, only two subjects that perform the HHI are present in the scene. In this paper, we assume the presence of more subjects besides the two involved in HHI in the scene, which fits better to the real scenarios. This significantly increases the chances of within-video mutual occlusions and needs more robust algorithms to identify the subjects involved in HHI.

In this paper, we propose a Graph Neural Network (GNN) based framework to jointly address the above-mentioned three tasks. Specifically, given two wearable-camera videos capturing a group of people with HHIs, we first model all the subjects in a video as a graph, in which each node represents a subject. We then propose a graph matching network by combining two videos to address the Task I of cross-view person identification. We further use a graph inference neural network to learn i) an adjacent matrix that represents the interactive relationship among all subjects (Task II) and ii) the node labels that represent the interaction category of each subject (Task III). Extensive experiments on the newly proposed video dataset validate the effectiveness of the proposed method.

This paper makes three major contributions: 1) This is the first work to propose and study the cross-view human identification and HHI detection in crowded multi-person scenes, which advances the human activity analysis to a more realistic scenario. 2) We study the mutual dependence and promotion of the above three tasks and propose a cascaded GNN to simultaneously achieve the graph matching and graph inference, implemented in an end-to-end way. The proposed Graph Inference Network (GIN) with dual-branch architecture can well integrate the information from two input videos. We jointly match the subjects, detect and recognize the interaction to form our multi-task model, while only a single task was addressed in previous works. 3) We construct a new multi-view video dataset and use it to evaluate the proposed method. We have released this new dataset to public[1].

2 RELATED WORK

2.1 Cross-View Person Identification

Cross-view person identification (CVPI) is a most fundamental problem in multi-view video analysis, which is to identify the same persons across the temporally synchronized videos taken by multiple (wearable) cameras [50]. Appearance is the most important and effective feature for object matching [31, 42, 43]. Motion feature can be also used for CVPI, however, its effectiveness is limited because of the view difference. Zheng et al. [50] extracted the view-invariant motion features by supervised deep learning and showed the effectiveness of using such features for CVPI. Liang et al. [20, 21] integrated the human pose features for CVPI by proposing a confidence-weighted human pose matching method, which can address the highly inaccurate 3D human pose estimation. Previous works only handle the scene containing one person without other pedestrians crossing through in the video [21, 50]. Differently, this paper aims to identify all the same persons across two videos with multiple people, which is very challenging due to the subjects' mutual occlusions and interactions.

[1] https://github.com/RuizeHan/CVID

2.2 Human-Human Interaction Recognition

Human activity detection is an important task in computer vision and multimedia computing. According to the number of participants, it can be divided into (single-) human action [18, 36], human-human interaction (HHI) and group event recognition/detection. Recently, a large proportion of works pay attention to the action and group event recognition. However, the study on modeling the interactions between two humans, i.e., HHI, is relatively fewer. Most of existing HHI methods and datasets focus on social or surveillance domains. Ryoo et al. [32] recognized the HHI in the videos taken from a fixed slope-angle top viewpoint with limited mutual occlusions. In [37], a dataset of human interaction clips with complementary pose data was introduced, where the videos are also captured from a fixed viewpoint and static background. Additional depth data is available in [46] for HHI recognition, in which the interactions fully occupy the frame. Some other researchers also focused on the HHI in TV shows [28], films [25], YouTube videos [14, 26, 49], and multi-view videos [27]. Most research on HHI has predominantly focused on *recognition rather than detection* [35], in which only the interactive humans appear in the video and the problem is to recognize their interaction types. Differently, we study a more realistic scene of detecting HHI in crowded multi-person scenes.

2.3 Multi-Person Video Analysis

Currently, most works on multi-person video analysis are focused on group event recognition, which aims to identify the ongoing activity involving a group of people. Classical graph models have been used for handling this problem [1–3, 16, 33, 40], resulting in many techniques, e.g., Markov Random Fields (MRFs) [40], Hierarchical Random Field (HiRF) [2] and spatial-temporal-aware graph [1, 33]. These methods model the individual actions and the group events by capturing the mutual relations among the people, e.g., human-level interactions [3], group-level interactions [1, 33] or multi-level interactions [16]. More recently, deep learning based models are widely applied for group event recognition [3, 6, 13, 23, 30, 38]. Among them, Recurrent Neural Networks (RNN) is widely used [6, 30, 38], while other networks like Hierarchical Relational Network [13] and Graph Convolutional Network (GCN) [23] have also been used for this problem. Note that, group event recognition identifies the overall activity of all the people, which is different from our problem of HHI detection in crowded scene. Also, all the previous works on group event recognition use a single fix-view video.

2.4 Graph Neural Network

Graph Neural Network (GNN) is an emerging research topic which has attracted extensive interest in the community [9, 41]. GNN integrates the advantages of classical graph models and popular neural networks with a strong relation representation and feature learning ability. GNN has been used in many tasks involving relation inference, such as human-object interaction (HOI) [7, 29], scene understanding [19, 24], human action localization [48] and human gaze communication [22]. GNN was also used to model the different parts of a human or other objects for action recognition [45] and object tracking [8]. More recently, several works [39, 47] were proposed to solve the deep graph matching problem, with verified the effectiveness in both theory and applications. Inspired by these efforts, we use the GNN to solve the proposed problem in this paper. More specifically, we employ a graph matching network to achieve the cross-view subject matching and a graph inference network to parse the interactive relation among the subjects.

3 OUR APPROACH

We design a unified graph neural network to jointly address the three tasks. Specifically, as shown in Figure 2, all the subjects are represented by graph nodes, and the HHI relations are represented by the adjacency matrix of graph. We use a Graph Matching Network (GMN) to match the subjects across the two views and a Graph Inference Network (GIN) to infer the interaction relation and category of each subject. We introduce the model formulation in Section 3.1 and the detailed network architecture in Section 3.2.

3.1 Model Formulation

Graph Representation. We first define a complete graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ to represent a multi-person scene at a frame. The graph node $v \in \mathcal{V}$ takes unique values from $\{1, \cdots, |\mathcal{V}|\}$, representing all the subjects in the frame. The graph edge $e = (v, w) \in \mathcal{E}$ indicates the connectivity of two nodes v, w, representing all the potential HHI relations. For node v, its node representation is denoted by a V-dimensional feature vector: $\mathbf{x}_v \in \mathbb{R}^{\mathrm{V}}$. Similarly, the edge representation for edge $e = (v, w)$ is denoted by an E-dimensional feature vector: $\mathbf{x}_{v,w} \in \mathbb{R}^{\mathrm{E}}$. Each graph node $v \in \mathcal{V}$ has an output state l_v that takes a value from a set of HHI type labels. We also define an adjacency matrix $\mathbf{A} \in \{0, 1\}^{|\mathcal{V}|\times|\mathcal{V}|}$ to represent the interaction relation over our complete graph $\mathcal{G}$, where each element $a_{v,w}$ represents the connectivity from node v to w.

Given a pair of synchronous frames in two videos capturing a scene from different views, we first construct two graphs $\mathcal{G}^1 = (\mathcal{V}^1, \mathcal{E}^1)$, $\mathcal{G}^2 = (\mathcal{V}^2, \mathcal{E}^2)$ as discussed above. As shown in Figure 2(b), we use a graph matching network to establish node-to-node correspondence between $\mathcal{G}^1$ and $\mathcal{G}^2$, which is denoted by a permutation matrix $\mathbf{P} \in \{0, 1\}^{|\mathcal{V}^1|\times|\mathcal{V}^2|}$ (Task I). Besides graph matching, for each graph $\mathcal{G}^{\mathrm{s}}(\mathrm{s} = 1, 2)$, our model also aims to learn the adjacency matrix $\mathbf{A}^{\mathrm{s}}$ and the interaction labels $\{l_v\}_{v\in\mathcal{V}^{\mathrm{s}}}$, i.e., the interaction types, of all the graph nodes $\mathcal{V}^{\mathrm{s}}$ simultaneously (Task II, III). In the following, we describe the above two steps.

Graph Matching Network. Given a pair of frames from different views, we first get the bounding box of each human (subject) by performing a human detector or using the ground-truth detections. As discussed above, each subject is represented by a graph node. We adopt a Convolutional Neural Network (CNN) for bounding box feature extraction, e.g., $\mathbf{f}_v$ denotes the initialized feature vector of v-th bounding box, where $v \in \mathcal{V}^{\mathrm{s}}(\mathrm{s} = 1, 2)$. As shown in Figure 2 (b), we use the intra-graph and cross-graph node embedding described in [39]. Specifically, the intra-graph node embedding is adopted as,

$$\tilde{\mathbf{x}}_v = \mathrm{GConv}(\tilde{\mathbf{x}}_v), \quad v \in \{\mathcal{V}^1, \mathcal{V}^2\}, \tag{1}$$

where $\mathrm{GConv} : \mathbb{R}^{\mathrm{V}} \rightarrow \mathbb{R}^{\mathrm{V}}$ denotes the graph convolution operation. For a pair of graphs $\mathcal{G}^1 = (\mathcal{V}^1, \mathcal{E}^1), \mathcal{G}^2 = (\mathcal{V}^2, \mathcal{E}^2)$, the cross-graph node embedding $\mathrm{CrossConv} : \mathbb{R}^{2\mathrm{V}} \rightarrow \mathbb{R}^{2\mathrm{V}}$ is adopted as,

$$\{\tilde{\mathbf{x}}_v, \tilde{\mathbf{x}}_u\} = \mathrm{CrossConv}(\tilde{\mathbf{x}}_v, \tilde{\mathbf{x}}_u), v \in \mathcal{V}^1, u \in \mathcal{V}^2. \tag{2}$$

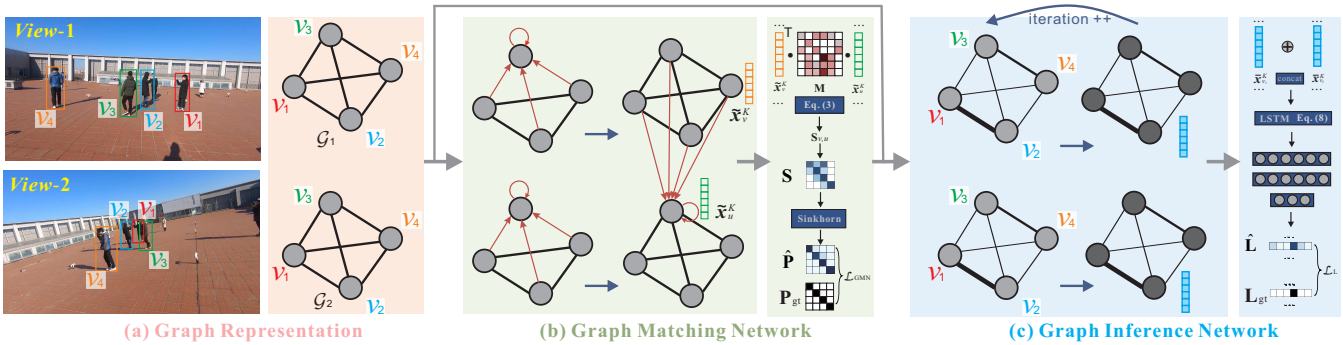

Figure 2: Framework of the proposed method. (a) A pair of synchronized frames from the given video pair and their corresponding graph representations. For simplicity, we only illustrate the graph structure of selective four subjects in each frame, where subjects v_1 and v_2 are under interaction of *waving*. (b) (c) Illustration of the proposed GMN and GIN.

As shown in Figure 2 (b), we take initial $\tilde{\mathbf{x}}_v$ as $\mathbf{f}_v$ and alternately adopt GConv, CrossConv, GConv for graph node embedding.

After getting the embedded node representation $\tilde{\mathbf{x}}_v, \tilde{\mathbf{x}}_u$ using the above embedding model, we consider calculating the affinity matrix $\mathbf{S} \in \mathbb{R}^{+^{|\mathcal{V}^1|\times|\mathcal{V}^2|}}$ containing the affinity score as

$$\mathbf{S}_{v,u} = \exp\left(\frac{\tilde{\mathbf{x}}_v^\top \mathbf{M} \tilde{\mathbf{x}}_u}{\tau}\right), \quad v \in \mathcal{V}^1, u \in \mathcal{V}^2, \tag{3}$$

where $\tilde{\mathbf{x}}_v, \tilde{\mathbf{x}}_u \in \mathbb{R}^{\mathrm{V}\times 1}$, and $\mathbf{M} \in \mathbb{R}^{\mathrm{V}\times\mathrm{V}}$ contains learnable weights to calculate the affinity matrix $\mathbf{S}$. $\tau > 0$ is a hyper-parameter to adjust the discriminative ability. Finally, we adopt the Sinkhorn operation to get the permutation matrix prediction $\hat{\mathbf{P}} = \mathrm{Sinkhorn}(\mathbf{S})$ [5]. Sinkhorn operation takes any non-negative square matrix and outputs a doubly-stochastic matrix, whose summation of each row or each column is one.

In the training process, the matrix cross-entropy loss is used for calculating the matching cost

$$\mathcal{L}_{\mathrm{GMN}} = \mathrm{L_P}(\mathbf{P}_{\mathrm{gt}}, \hat{\mathbf{P}}), \tag{4}$$

where $\mathbf{P}_{\mathrm{gt}}$ and $\hat{\mathbf{P}}$ are the ground-truth and predicted permutation matrix. In the testing process, we apply the Hungarian algorithm [15] on permutation matrix prediction $\hat{\mathbf{P}}$ as a post processing step to discretize the output into a binary permutation matrix $\mathbf{P}$.

Graph Inference Network. For simplicity, we first describe the graph inference network (GIN) in terms of one branch. Given a graph $\mathcal{V} = \mathcal{V}^1$ (or $\mathcal{V}^2$), the GIN updates the adjacency matrix $\mathbf{A}$ to infer the current interaction graph structure, according to the node and edge representations

$$a_{v,w}^{(k)} = \sigma(\mathrm{F_A}(\bar{\mathbf{x}}_v^{(k-1)}, \bar{\mathbf{x}}_w^{(k-1)}, \bar{\mathbf{x}}_{v,w}^{(k-1)})), v, w \in \mathcal{V}, \tag{5}$$

where the connectivity matrix $\mathbf{A}^{(k)} = [a_{v,w}^{(k)}]_{v,w} \in \mathbb{R}^{|\mathcal{V}|\times|\mathcal{V}|}$ encodes current interaction relation predictions. $\mathrm{F_A} : \mathbb{R}^{2\mathrm{V}+\mathrm{E}} \to \mathbb{R}^1$ is a connectivity readout network that maps an edge representation into the connectivity weight and σ is an activation function. Besides, in the node update phase, we update node representations $\bar{\mathbf{x}}_v$ via considering all the incoming node and edge information weighted by the corresponding connectivity

$$\bar{\mathbf{x}}_v^{(k)} = \sigma\Big(\sum\nolimits_w a_{v,w}^{(k-1)} \mathrm{F_V}(\bar{\mathbf{x}}_v^{(k-1)}, \bar{\mathbf{x}}_w^{(k-1)}, \bar{\mathbf{x}}_{v,w}^{(k-1)})\Big), \tag{6}$$

where $\mathrm{F_V} : \mathbb{R}^{2\mathrm{V}+\mathrm{E}} \to \mathbb{R}^{\mathrm{V}}$ represents a node update network. As shown in Figure 2(c), we take $\bar{\mathbf{x}}_v^{(0)} = \mathbf{f}_v$ as the initial node representation. We iteratively update the adjacency matrix and node representation for K iterations then obtain $\hat{\mathbf{A}} = \mathbf{A}^{(K)}$ and $\bar{\mathbf{x}}_v^{(K)}$.

Next, as shown in Figure 2(c), given the cross-view subject matching results, we concat the node features representing the same subject in two views and get the syncretic representations $\mathbf{y}_v$

$$\mathbf{y}_v = \mathrm{concat}(\bar{\mathbf{x}}_{v_1}^{(K)}, \bar{\mathbf{x}}_{v_2}^{(K)}), v_1 \in \mathcal{V}^1, v_2 \in \mathcal{V}^2, \tag{7}$$

where v_1, v_2 denote the same subject v appearing in two views, and the cross-view subject matching results are obtained by GMN. We then use an LSTM function layer to update the node representation by considering temporal information.

$$\mathbf{h}_v^t = \mathrm{F_{LSTM}}(\mathbf{y}_v^t | \mathbf{h}_v^{t-1}), \tag{8}$$

where $\mathbf{y}_v^t$ denotes the input of the LSTM at frame t, and $\mathbf{h}_v^t$ denotes the corresponding hidden state considering the previous information $\mathbf{h}_v^{t-1}$. Finally, we use a readout function $\mathrm{F_R}(\cdot)$ followed by an activation function σ to output the interaction labels of each subject

$$\hat{l}_v = \sigma(\mathrm{F_R}(\mathbf{h}_v^t)), \tag{9}$$

from which we can obtain the node label matrix $\hat{\mathbf{L}} = [\hat{l}_v]_v \in \mathbb{R}^{|\mathcal{V}|}$.

For GIN, we define the following loss

$$\mathcal{L}_{\mathrm{GIN}} = \sum_{s=1,2} \mathrm{L_A}(\mathbf{A}_{\mathrm{gt}}^s, \hat{\mathbf{A}}^s) + \mathrm{L_L}(\mathbf{L}_{\mathrm{gt}}, \hat{\mathbf{L}}), \tag{10}$$

where the loss functions $\mathrm{L_A}$ and $\mathrm{L_L}$ calculate the distance between the predicted and ground-truth adjacent matrix/node labels. $\hat{\mathbf{A}}^s$ and $\mathbf{A}_{\mathrm{gt}}^s (s = 1, 2)$ denote the predicted and ground-truth adjacency matrix in two branches and $\hat{\mathbf{L}}$ and $\mathbf{L}_{\mathrm{gt}}$ denote the node labels.

We provide an example to intuitively illustrate the proposed method as shown in Figure 2. Given a pair of synchronized video frames, we represent all the subjects in each frame as a complete graph by taking each subject as node and the relation between two subjects as edge. Then the initial graphs are first imported into the GMN 2(b). The network iteratively updates the graph node representation by intra-graph and inter-graph embeddings using Eq. (1) and Eq. (2), respectively, as shown in the left of Figure 2(b). After that, the similarity of every two nodes from different graphs is calculated by Eq. (3) to compose the affinity matrix. Through the

Sinkhorn operation, the predicted permutation matrix is used to compute the cross-entropy loss against the ground-truth permutation matrix using Eq. (4). The matching results is used for the following GIN 2(c), which also takes the initial graphs as input. As shown in the left of Figure 2(c), the networks iteratively update the adjacency matrix (note the change of edge thickness) and node representation (note the change of node gray-levels) using Eq. (5) and Eq. (6), respectively. Next, we combine the updated node representations of the same subject from different videos (using the matching results) for the interaction category recognition in the right of (c). For this purpose, we use an LSTM function layer Eq. (8) to capture the temporal information followed by a fully connected layer Eq. (9) to predict the final interaction labels. The proposed method can be summarized in Algorithm 1.

Algorithm 1: Human identification and HHI detection:

Input: $\mathcal{G}^s = (\mathcal{V}^s, \mathcal{E}^s)$ (s = 1,2); pre-set parameters.
Output: The permutation matrix $\hat{\mathbf{P}}$, the adjacency matrix $\mathbf{A}^s$ and the interaction labels $\{l_v\}_{v \in \mathcal{V}^s}$ (s = 1,2).

```
Divide the video segment into clips of T frames.
for t = 1 : T do
    Build the two graphs G^1 and G^2 on frame t of the two clips.
    Get the node embedding for all nodes in G^1 and G^2 using
      GConv, CrossConv, GConv in Eqs. (1, 2) alternately.
    Calculte the affinity matrix by Eq. (3) and learn M.
    Adopt the Sinkhorn operation to get P̂.
    for s = 1 : 2 do
        for k = 1 : K do
            Update the adjacency matrix A_s^(k) by Eq. (5).
            Update node representations x̄_s^(k) by Eq. (6).
    Connect the matched x̄_s^(K) by Eq. (7) using P̂.
    Update the node representation by the LSTM function of Eq. (8).
    Output the interaction labels by the readout function of Eq. (9).
```

3.2 Detailed Network Architecture

Node and Edge Representation. We use the annotated detection result of each subject. For each node $v \in \mathcal{V}$, the initial feature $\mathbf{f}_v$ combines 1) appearance feature: extracted from the corresponding bounding box using a pre-trained person re-identification (Re-ID) network [51]), 2) pose feature: extracted from the corresponding human with an existing pose estimation approach [4], and 3) location feature: a 6-d subject position information, i.e., the coordinates of the up-left corner, the center points and the bottom-right corner of the corresponding bounding box. In graph matching network, we directly use the appearance feature as the node representation. In graph inference network, we combine the pose feature and location feature because the appearance feature is not very useful for human interaction detection and recognition. Moreover, similar to [29, 39], to decrease the amount of parameters and make the dimensions of different features comparable, we use a post processing to compress the pose features into 8-d and combine it to the 6-d location feature as initial node feature. For an edge $e \in \mathcal{E}$, it is represented by combining the features of the two linked nodes.

Loss Functions. The matrix cross-entropy loss between two matrices used in Eq. (4) is defined as

$$\mathrm{L}(\mathbf{X}, \mathbf{Y}) = -\mathbf{1}^\top (\mathbf{X} \odot \log \mathbf{Y} + (1 - \mathbf{X}) \odot \log(1 - \mathbf{Y}))\mathbf{1} \tag{11}$$

where we assume $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{K_1 \times K_2}$, $\mathbf{1}^\top \in \mathbb{R}^{1 \times K_1}$, $\mathbf{1} \in \mathbb{R}^{K_2 \times 1}$, and $\odot$ denotes the element-wise multiplication operation. For the loss of graph inference network in Eq. (10), we use L_1 loss function for L_A and the cross-entropy loss for L_L. The total loss function of the whole framework is defined as

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\mathrm{GMN}} + \lambda_2 \mathcal{L}_{\mathrm{GIN}}, \tag{12}$$

where $\mathcal{L}_{\mathrm{GMN}}$ and $\mathcal{L}_{\mathrm{GIN}}$ are the graph matching network loss and the graph inference network loss, respectively. We use two parameters λ_1 and λ_2 as weight coefficients.

Graph Network Implementations. The adjacency matrix update functions $\mathrm{F_A}$ in Eq. (5) is implemented by a four-layer convolutional network. The node update function $\mathrm{F_V}$ in Eq. (6) is implemented by the fully-connected-layer and gated recurrent unit (GRU) network. We use a single-layer BiLSTM as the function $\mathrm{F_{LSTM}}$ in Eq. (8). The readout function $\mathrm{F_R}$ in Eq. (9) is implemented by the fully-connected-layer network. Besides, the activity function σ in Eq. (5) and (6) is the sigmoid function, and in (9) is ReLU function. Our model is implemented by PyTorch on an NVIDIA GTX-2080Ti GPU. During training phase, the learning rate is set to 1×10^{-3}, and decays to 0.1 times every 10 epochs. The training process takes 20 epochs to converge with a batch size of 16. The parameters τ in Eq. (3) is set to 5×10^{-3}. The parameters λ_1 and λ_2 in Eq. (12) are set to 1 and 0.5, respectively. The iterations K in GIN is set to 2 in our experiment. The training/testing dataset splitting will be discussed in Section 4.1. For training the network, to alleviate the unbalance of training samples, we only sample the interval from the starting frame to the ending frame of an interaction. Each video pair is imported into the network clip by clip with fixed length e.g., 8 frames, in our experiments. The GConv and CrossConv in GMN and the functions of Eqs. (5, 6, 8, 9) in GIN are implemented by neural networks, whose parameters are learnable. The proposed GMN and GIN are trained in an end-to-end way.

Temporal Smoothness. Although the GNN based framework discussed above has considered the temporal information of the sequence, the frame-wise interaction detection and recognition output may provide noisy predictions for some frames. For example, the network may wrongly predict the interaction category for a small number of frames in the interval of an interaction, or predict an interaction during the interval without any interaction. To reduce noise and preserve the prediction continuity along a sequence, we design a simple yet effective long-short-term sliding window based approach for temporal smoothness. A short-term sliding window, e.g., 6 frames, is used to correct the error predictions on single frames. When an interaction label prediction on one frame is different from the others in this window, it will be corrected. Similarly, we also use a long-term sliding window, e.g., 30 frames, to correct the error predictions on the short clips. Using this strategy, the temporal interaction detection and recognition can be more continuous between frames and therefore, more accurate. We will show the effectiveness of this post processing in the experiment section.

Figure 3: Example frames of our dataset. Same human objects across two views are marked by boxes with the same color.

4 EXPERIMENTS

4.1 Dataset and Metrics

Data Collection. We do not find publicly available datasets with multi-view multi-subject videos with ground-truth annotations for HHI detections. Therefore, in this work we collect a new video dataset using multiple GoPro wearable cameras for performance evaluation. This dataset contains five common categories of social HHIs: (hand) shaking, hugging, exchanging, waving, and patting. We invite 10 subjects to randomly walk or stand in the scene, and two of them wear GoPro cameras overhead to collect videos. The locations of the two cameras are always separated and their view directions differ by more than 90 degrees, which ensures the two collected videos to provide complementary information for subject/activity detection and recognition. Given large view difference, the background in the two videos can be totally different. The videos are recorded in a way that two selected subjects (not the camera wearers) perform one of the five interactions at a time, while all the other subjects are free to move or stop, resulting random mutual occlusions in these videos. These videos also contain no-interaction intervals (with random lengths) between two HHIs. We *manually synchronize* the pair of videos taken by the two GoPros such that corresponding frames between them are taken at the same time. For dataset annotation, we hire three expert volunteers to independently annotate the starting and ending frames of each HHI and then take their averages as the ground truth interval of the HHI detection. The corresponding involved subjects as well as the interaction category of each HHI are also annotated. The bounding boxes of all the subjects are also labeled by outsourcing to a professional company. Example frames and our ground-truth annotations are shown in Figure 3, from which we can see that these paired wearable-camera videos do provide complementary information for HHI analysis. We can also see that the frequent mutual occlusions and perspective difference bring challenges to our tasks.

Data Statistics. In total, we collect 75 pairs of videos with 48,600 frames in our dataset. Specifically, as shown in Table 1, the collected 75 pairs of videos contain the above-mentioned five categories of HHIs. Each pair of videos contain two *randomly selected* HHIs (with same or different categories) occurred over non-overlapped time intervals. There exist random-length no-interaction intervals before the first HHI, after the second HHI and between the two HHIs. The second column of Table 1 shows the total number of frames in these videos. We annotate each frame with bounding boxes of all the subjects and 367,191 human bounding boxes (avg. 7.6 per frame) are annotated in total. We also annotate the interaction interval (the starting and ending frames), as well as the involved subjects and corresponding interaction labels frame by frame, resulting in 11,700 HHI labels (frame level) in total. Note, the data volume for different HHI categories are balanced. As shown in the last three rows of Table 1, we split the dataset into training and testing data by 3:2 with no overlap.

Table 1: Dataset statistics and training/testing splitting.

Dataset	# Video pair	# Frame	# Subject	# Inter.
shaking	15	8,600	64,194	2,330
hugging	15	10,310	77,535	2,640
exchange	15	10,750	84,525	2,210
waving	15	8,870	66,306	2,390
patting	15	10,070	74,631	2,130
training	45	31,030	233,847	7,180
testing	30	17,570	133,344	4,520
full	75	48,600	367,191	11,700

Evaluation Metrics. Task I: We use precision ($\mathcal{P}$), recall ($\mathcal{R}$) and F_1-score ($\mathcal{F}$) for evaluation. Precision $\mathcal{P}$ / recall $\mathcal{R}$ denote the ratio of true-positive matches to all predicted-positive / real-positive matches, respectively. F_1 score $\mathcal{F}$ is computed as $\mathcal{F} = \frac{2\times\mathcal{P}\times\mathcal{R}}{\mathcal{P}+\mathcal{R}}$.

Task II : 1) Temporal domain – A frame with identical predicted and ground-truth interaction label is taken as true-positive detection. We compute the ratio of true-positive detections to the numbers of frames with predicted and ground-truth interactions and get the temporal-domain precision $\mathcal{P}$ and recall $\mathcal{R}$, respectively. We then compute F_1-score ($\mathcal{F}$) based on $\mathcal{P}$ and $\mathcal{R}$. We also use average accuracy ($\mathcal{A}$) by averaging the accuracy of temporal interaction detection for all frames, including the frames with and without interaction. **2) Spatial domain** – On each frame, we use an adjacency matrix to represent the interactive relations among all the subjects. For a predicted adjacency matrix $\hat{\mathbf{A}} \in \mathbb{R}^{N\times N}$ and a ground-truth adjacency matrix $\mathbf{A}_{\text{gt}} \in \mathbb{R}^{N\times N}$, where N denotes the number of subjects. Then the spatial-domain precision $\mathcal{P}$ and recall $\mathcal{R}$ are $\mathcal{P} = \frac{\sum \text{AND}(\hat{\mathbf{A}}, \mathbf{A}_{\text{gt}})}{\sum \hat{\mathbf{A}}}, \mathcal{R} = \frac{\sum \text{AND}(\hat{\mathbf{A}}, \mathbf{A}_{\text{gt}})}{\sum \mathbf{A}_{\text{gt}}}$, where AND denotes the logical function, the numerator counts true positive HHI relations, and the denominators count the predicted and ground-truth HHI relations, respectively.

Task III: We compute the Top-1 and Top-2 classification accuracy ($\mathcal{T}$-1 and $\mathcal{T}$-2) to evaluate the interaction category recognition (classification) performance for the subjects with correct

spatial-temporal interaction detection. We also propose a new comprehensive metric – multiple human interaction accuracy (MHIA) for Task II and Task III. Specifically, following the standard metric MOTA (multiple object tracking accuracy) in MOT problem [17], MHIA is computed as $\text{MHIA} = 1 - \frac{\sum_t(\text{ms}_t+\text{fp}_t+\text{fc}_t)}{\sum_t(\text{d}_t+\text{g}_t)}$, where ms_t, fp_t are the numbers of missed (false negative), false positive subjects for spatial-domain interaction detection at time t, fc_t denotes the number of subjects with true interaction detection but false interaction category, and d_t and g_t represent the number of detected and ground-truth subjects with interaction at time t, respectively.

4.2 Baselines

We consider following baselines for Task I:

• *Chance*: Randomly create the permutation matrix for matching.

• VGG + Hungary/DHN: Compute the affinity matrix by the similarity between the feature vectors extracted by VGG network [34]. Adopt the Hungary algorithm [15] / Deep Hungary Network (DHN) [44] to get the permutation matrix as the matching results.

• Re-ID + Hungary / DHN: Similar to the above methods and use the appearance feature extracted by the pre-trained re-ID network [51].

• DL-GM / PCA-GM: Build the graph as in Section 3.1 and use a deep learning based Graph Matching (GM) method proposed in [47] and [39], respectively. Note, the training/testing data for DHN, DL-GM and PCA-GM are the same as in our method.

For Tasks II, III, we do not find directly related comparative methods. Prior HHI detection methods focus on the scene only containing the interactive persons, which can not handle the scene with more subjects other than the interactive ones, and identify the interactive subjects. Differently, we found that some human-object interaction detection methods can identify the interactive relations among different humans/objects. We consider the baselines:

• *Chance*: A weak baseline that randomly assigns an interaction label to each subject node.

• *CNN + LSTM*: An alternative method that uses five-layer Conv2d network following by an LSTM for label classification, where it only considers the temporal dynamics but no spatial structures and predicts the subject's activity individually.

• *GPNN* [29]: A human-object interaction detection method using the GNN, which is re-trained on the proposed dataset.

4.3 Cross-View Human Identification Results

As shown in Table 2, we first evaluate the CVPI (Task I). Our full model achieves the best performance with an F_1 score of 76.5%. Compared to the Hungary and DHN algorithm based approaches, the proposed method performs better when using the same appearance feature from [51]. This comparison demonstrates that the proposed graph based network can provide stronger representation and discrimination ability for CVPI. Moreover, we can also see that our approach gets better performance than the baseline deep graph matching method DL-GM when using the same features extraction method and training data. We further evaluate the matching performance of our method without the graph inference network, i.e., 'w/o GIN' in Table 2. We can see that the graph matching network only performs worse than the whole framework. It can be explained that the combination of solving human matching and interaction detection promotes the human identification task.

Table 2: Results of cross-view person identification (Task I).

Method	$\mathcal{P}$ (%)	$\mathcal{R}$ (%)	$\mathcal{F}$ (%)
Chance	10.1	14.3	11.8
VGG + Hungary	13.8	18.3	15.7
VGG + DHN [44]	14.2	20.1	16.6
Re-ID [51] + Hungary	57.1	76.0	65.2
Re-ID [51] + DHN [44]	56.7	80.5	66.5
DL-GM [47]	44.1	58.7	50.3
w/o GIN	63.5	84.5	72.5
Ours	**66.9**	**89.3**	**76.5**

4.4 HHI Detection and Recognition Results

As shown in Table 3, we can see that all the baseline methods produce poor results in both Task II and Task III, due to the originality and challenging of our problem. Our full model achieves the best performance which leads a large margin compared to the baselines.

Ablation Study. We derive the following variants of our method to evaluate the effectiveness of our essential components:

• *w single view:* uses the single-view video data for training GIN without cross-view data fusion strategy.

• *w/o pose/location:* removes the human pose/location feature.

• *w/o temporal:* removes the LSTM function F_{LSTM} in Eq. (8) for temporal information modeling.

• *w/o **A**-supervision:* removes the ground-truth adjacent matrix **A** by unsupervisedly learning it.

• *w iter-1/3*: changes the iteration time K for graph learning in GIN from 2 into 1 and 3, respectively.

• *w/o smoothness:* removes the temporal smoothness strategy.

• *w **P**-GT:* uses the annotated subject matching results in GIN.

For in-depth analysis, we can see that our model using *single-view* video performs not very well, which is because the mutual occlusion and background clutter in the multi-person scene disturb the detection and recognition accuracy. We can also see that both the *pose* and *location* features are effective in our model. Note that, *w/o location* provides unexpected recognition results. This is because it detects very few interactive subjects, which can be seen from the recall score of temporal-spatial-domain interaction detection results. We then study the architecture design of GIN. The method *w/o **A**-supervision* which unsupervisedly learns the adjacency matrix without using the ground-truth **A** obtains poor results. This is because that interaction detection is not simply about individual feature (node representation), but also dependent on a comprehensive inference of spatial-temporal relations (adjacency matrix). We can also see that the *temporal* information obtained by the LSTM function is effective to a certain extent. We also find that the iteration time K for graph learning in GIN performs best when setting to 2 . However, the performance is not very sensitive to the iteration number. We examine the effect of temporal *smoothness* strategy as post-processing and find it is able to gradually improve the performance in general. To independently evaluate the effectiveness of GIN, we adopt the annotated CVPI results (ground truth)

Table 3: Comparative results of spatial-temporal-domain HHI detection (Task II) and recognition (Task III).

Method	Temporal Domain				Spatial Domain			Recognition		Overall
	$\mathcal{A}$ (%)	$\mathcal{P}$ (%)	$\mathcal{R}$ (%)	$\mathcal{F}$ (%)	$\mathcal{P}$ (%)	$\mathcal{R}$ (%)	$\mathcal{F}$ (%)	$\mathcal{T}$-1(%)	$\mathcal{T}$-2(%)	MHIA (%)
Chance	50.0	44.0	48.7	46.2	2.7	3.0	2.9	13.5	35.1	7.0
CNN + LSTM	56.7	53.3	41.5	46.7	31.3	24.4	27.4	53.4	60.1	27.9
GPNN [29]	45.0	58.4	69.5	63.5	48.5	57.7	52.7	24.0	41.0	37.1
w single-view	45.4	46.0	93.6	61.6	36.5	74.9	49.0	34.0	47.3	36.6
w/o pose	44.1	42.3	70.6	52.9	32.3	53.9	40.4	15.0	28.4	27.1
w/o location	62.0	77.6	14.5	24.5	77.2	14.4	24.3	57.5	66.2	19.2
w/o **A**-supervision	61.4	83.7	13.8	23.7	55.4	9.2	15.7	37.3	46.7	11.7
w/o temporal	48.0	51.1	91.1	65.5	44.2	78.8	56.7	36.7	54.2	40.9
w iter-1	71.9	71.3	48.0	57.4	68.5	46.1	55.1	50.8	62.9	42.3
w iter-3	54.8	53.1	71.4	60.9	47.9	64.4	54.9	44.5	67.0	41.6
w/o smoothness	64.3	57.3	76.1	65.4	51.3	68.2	58.6	46.1	57.3	45.0
w **P**-GT	65.7	59.9	79.6	68.3	54.3	72.1	61.9	51.2	58.3	49.1
Ours (w iter-2)	65.7	58.8	71.8	64.6	53.8	65.8	59.2	50.9	58.9	46.7

for GIN. This way, *w **P**-GT* naturally gets the better performance, which is also reasonable.

Qualitative Results. We show sample visual results of our full method for HHI detection and recognition in Figure 4, where the subjects with/without interaction are in green/gray bounding boxes while the predicted interaction category labels (green for predicted label and red for ground-truth label) are shown beside the boxes. We can see that the proposed method can largely localize the HHI occurrence interval while some inaccuracies only happen around the starting and ending time of each HHI. For the temporal-domain interaction detection, we show the frame-by-frame detection results in Figure 4. For the spatial-domain interaction detection, from the second and fourth columns at the top of Figure 4, we can see that our method can localize the HHI subjects in the crowded scene with many people. Also, the interaction category is correctly recognized as shown in the labels beside the bounding boxes. For further analysis, in some cases, the interaction is clear in one view but indistinct in the other, e.g., the second column in Figure 4. For more serious cases, one subject with interaction is fully occluded by the others and can not be seen in one view, as shown in the fourth column. By this way, the cross-view videos can provide complementary information for interaction detection and recognition. Figure 5(a) shows a failure case. We can see that the true interactive subjects are 'A' and 'C', while the predicted are 'A' and 'B'. This can be explained that the false-positive interaction prediction confidence of 'A' and 'B' is very high in 'View-1'. In this case, it is difficult to get correct prediction given that the underlying interactive activities are fully/partially occluded in both views. To address such problems, we may need more cameras with different views. Figure 5(b) shows a failure case of interaction category recognition, where the predicted category 'exchanging' looks very similar to the ground-truth category 'shaking'.

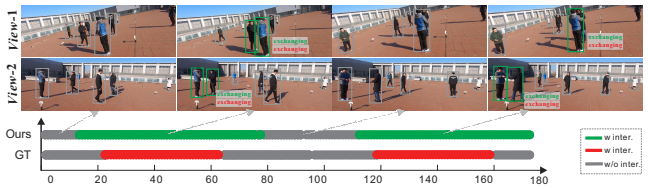

Figure 4: An illustration of qualitative results.

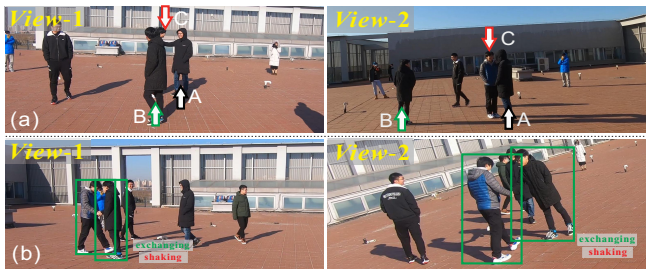

Figure 5: An illustration of failure cases.

5 CONCLUSION

This paper proposed and addressed a new problem of cross-view human identification and interaction detection in multi-person scene captured by two wearable cameras. We simultaneously studied the mutual dependence and promotion of three important tasks in video surveillanc. We propose a GNN based framework to address our problem by exploiting the advantages of graph structure in modeling the relations among multiple subjects. We also collected a new video dataset to evaluate the proposed approach and the results verified the effectiveness of our method. Through the above efforts, we just hope to provide the resources for studying this new problem and move the human activity analysis one more step toward realistic scenarios. In the future, we plan to study the human interaction detection and recognition in the scenario where multiple interactive activities occur simultaneously.

Acknowledgement. This work was supported in part by the NSFC under Grants U1803264, 61672376, 61572354, 61671325, and by the research fund for the Tianjin Key Lab for Advanced Signal Processing, Civil Aviation University of China, No.2019ASP-TJ01.

REFERENCES

[1] Mohamed Amer, Michael J. Black, and Ivan Laptev. 2012. Cost-Sensitive Top-Down/Bottom-Up Inference for Multiscale Activity Recognition. In *ECCV*.

[2] Mohamed R Amer, Peng Lei, and Sinisa Todorovic. 2014. HiRF: Hierarchical Random Field for Collective Activity Recognition in Videos. In *ECCV*.

[3] Timur Bagautdinov, Alexandre Alahi, François Fleuret, Pascal Fua, and Silvio Savarese. 2017. Social Scene Understanding: End-to-End Multi-Person Action Localization and Collective Activity Recognition. In *CVPR*.

[4] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2017. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*.

[5] Rodrigo Santa Cruz, Basura Fernando, Anoop Cherian, and Stephen Gould. 2019. Visual Permutation Learning. *IEEE TPAMI* 41, 12 (2019), 3100–3114.

[6] Zhiwei Deng, Arash Vahdat, Hexiang Hu, and Greg Mori. 2016. Structure Inference Machines: Recurrent Neural Networks for Analyzing Relations in Group Activity Recognition. In *CVPR*.

[7] Hao-Shu Fang, Jinkun Cao, Yu-Wing Tai, and Cewu Lu. 2018. Pairwise body-part attention for recognizing human-object interactions. In *ECCV*.

[8] Junyu Gao, Tianzhu Zhang, and Changsheng Xu. 2019. Graph Convolutional Tracking. In *CVPR*.

[9] Xiang Gao, Wei Hu, Jiaxiang Tang, Jiaying Liu, and Zongming Guo. 2019. Optimized skeleton-based action recognition via sparsified graph regression. In *ACM MM*.

[10] Ruize Han, Wei Feng, Jiewen Zhao, Zicheng Niu, Yujun Zhang, Liang Wan, and Song Wang. 2020. Complementary-View Multiple Human Tracking. In *AAAI*.

[11] Ruize Han, Yujun Zhang, Wei Feng, Chenxing Gong, Xiaoyu Zhang, Jiewen Zhao, Liang Wan, and Song Wang. 2019. Multiple Human Association between Top and Horizontal Views by Matching Subjects' Spatial Distributions. *arXiv preprint arXiv:1907.11458* (2019).

[12] Ruize Han, Jiewen Zhao, Wei Feng, Yiyang Gan, Liang Wan, and Song Wang. 2020. Complementary-View Co-Interest Person Detection. In *ACM MM*.

[13] Mostafa S Ibrahim and Greg Mori. 2018. Hierarchical Relational Networks for Group Activity Recognition and Retrieval. In *ECCV*.

[14] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. 2017. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950* (2017).

[15] Harold W Kuhn. 1955. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly* 2, 1 (1955), 83–97.

[16] Tian Lan. 2012. Discriminative Latent Models for Recognizing Contextual Group Activities. *IEEE TPAMI* 34, 8 (2012), 1549–1562.

[17] Laura Leal-Taixé, Anton Milan, Ian Reid, Stefan Roth, and Konrad Schindler. 2015. Motchallenge 2015: Towards a benchmark for multi-target tracking. *arXiv preprint arXiv:1504.01942* (2015).

[18] Qing Li, Zhaofan Qiu, Ting Yao, Tao Mei, Yong Rui, and Jiebo Luo. 2016. Action recognition by learning deep multi-granular spatio-temporal video representation. In *ACM MM*.

[19] Ruiyu Li, Makarand Tapaswi, Renjie Liao, Jiaya Jia, Raquel Urtasun, and Sanja Fidler. 2017. Situation Recognition with Graph Neural Networks. In *ICCV*.

[20] Guoqiang Liang, Xuguang Lan, Xingyu Chen, Kang Zheng, Song Wang, and Nanning Zheng. 2019. Cross-View Person Identification Based on Confidence-Weighted Human Pose Matching. *IEEE TIP* 28, 8 (2019), 3821–3835.

[21] Guoqiang Liang, Xuguang Lan, Kang Zheng, Song Wang, and Nanning Zheng. 2018. Cross-View Person Identification by Matching Human Poses Estimated with Confidence on Each Body Joint. In *AAAI*.

[22] Siyuan Huang Xinyu Tang Song-Chun Zhu Lifeng Fan, Wenguan Wang. 2019. Understanding Human Gaze Communication by Spatio-Temporal Graph Reasoning. In *CVPR*.

[23] Lihua Lu, Yao Lu, Ruizhe Yu, Huijun Di, Lin Zhang, and Shunzhou Wang. 2020. GAIM: Graph Attention Interaction Model for Collective Activity Recognition. *IEEE TMM* 22, 2 (2020), 524–539.

[24] Kenneth Marino, Ruslan Salakhutdinov, and Abhinav Gupta. 2016. The more you know: Using knowledge graphs for image classification. *arXiv preprint arXiv:1612.04844* (2016).

[25] Marcin Marszalek, Ivan Laptev, and Cordelia Schmid. 2009. Actions in context. In *CVPR*.

[26] Mathew Monfort, Alex Andonian, Bolei Zhou, Kandan Ramakrishnan, Sarah Adel Bargal, Tom Yan, Lisa Brown, Quanfu Fan, Dan Gutfreund, Carl Vondrick, et al. 2019. Moments in Time Dataset: one million videos for event understanding. *IEEE TPAMI* 42, 2 (2019), 502–508.

[27] Saeid Motiian, Farzad Siyahjani, Ranya Almohsen, and Gianfranco Doretto. 2017. Online Human Interaction Detection and Recognition With Multiple Cameras. *IEEE TCSVT* 27, 3 (2017), 649–663.

[28] Alonso Patron-Perez, Marcin Marszalek, Ian Reid, and Andrew Zisserman. 2012. Structured Learning of Human Interactions in TV Shows. *IEEE TPAMI* 34, 12 (2012), 2441–2453.

[29] Siyuan Qi, Wenguan Wang, Baoxiong Jia, Jianbing Shen, and Songchun Zhu. 2018. Learning Human-Object Interactions by Graph Parsing Neural Networks. In *ICCV*.

[30] Vignesh Ramanathan, Jonathan Huang, Sami Abu-El-Haija, Alexander Gorban, Kevin Murphy, and Li Fei-Fei. 2016. Detecting Events and Key Actors in Multi-person Videos. In *CVPR*.

[31] Ergys Ristani and Carlo Tomasi. 2018. Features for Multi-target Multi-camera Tracking and Re-identification. In *CVPR*.

[32] Michael S Ryoo and Jake K Aggarwal. 2009. Spatio-Temporal Relationship Match: Video Structure Comparison for Recognition of Complex Human Activities. In *ICCV*.

[33] Tianmin Shu, Dan Xie, Brandon Rothrock, Sinisa Todorovic, and Songchun Zhu. 2015. Joint inference of groups, events and human roles in aerial videos. In *CVPR*.

[34] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. In *CVPR*.

[35] Alexandros Stergiou and Ronald Poppe. 2019. Analyzing human–human interactions: A survey. *CVIU* 188 (2019), 102799.

[36] Yi Tian, Qiuqi Ruan, Gaoyun An, and Yun Fu. 2016. Action recognition using local consistent group sparse coding with spatio-temporal structure. In *ACM MM*.

[37] Coert Van Gemeren, Ronald Poppe, and Remco C Veltkamp. 2016. Spatio-Temporal Detection of Fine-Grained Dyadic Human Interactions. In *International Workshop on Human Behavior Understanding*.

[38] Minsi Wang, Bingbing Ni, and Xiaokang Yang. 2017. Recurrent Modeling of Interaction Context for Collective Activity Recognition. In *CVPR*.

[39] Runzhong Wang, Junchi Yan, and Xiaokang Yang. 2019. Learning Combinatorial Embedding Networks for Deep Graph Matching. In *ICCV*.

[40] Zhenhua Wang, Qinfeng Shi, Chunhua Shen, and Anton Van Den Hengel. 2013. Bilinear Programming for Human Activity Recognition with Unknown MRF Graphs. In *CVPR*.

[41] Jiaxin Wu, Sheng-Hua Zhong, and Yan Liu. 2019. MvsGCN: A Novel Graph Convolutional Network for Multi-video Summarization. In *ACM MM*.

[42] Yuanlu Xu, Xiaobai Liu, Yang Liu, and Song-Chun Zhu. 2016. Multi-view People Tracking via Hierarchical Trajectory Composition. In *CVPR*.

[43] Yuanlu Xu, Xiaobai Liu, Lei Qin, and Song-Chun Zhu. 2017. Cross-View People Tracking by Scene-Centered Spatio-Temporal Parsing. In *AAAI*.

[44] Yihong Xu, Aljosa Osep, Yutong Ban, Radu Horaud, Laura Lealtaixe, and Xavier Alamedapineda. 2019. How To Train Your Deep Multi-Object Tracker. In *CVPR*.

[45] Sijie Yan, Yuanjun Xiong, and Dahua Lin. 2018. Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition. In *AAAI*.

[46] Kiwon Yun, Jean Honorio, Debaleena Chattopadhyay, Tamara L. Berg, and Dimitris Samaras. 2012. Two-person interaction detection using body-pose features and multiple instance learning. In *CVPRW*.

[47] Andrei Zanfir and Cristian Sminchisescu. 2018. Deep Learning of Graph Matching. In *CVPR*.

[48] Runhao Zeng, Wenbing Huang, Chuang Gan, Mingkui Tan, Yu Rong, Peilin Zhao, and Junzhou Huang. 2019. Graph Convolutional Networks for Temporal Action Localization. In *ICCV*.

[49] Hang Zhao, Antonio Torralba, Lorenzo Torresani, and Zhicheng Yan. 2019. HACS: Human Action Clips and Segments Dataset for Recognition and Temporal Localization. In *ICCV*.

[50] Kang Zheng, Xiaochuan Fan, Yuewei Lin, Hao Guo, and Song Wang. 2017. Learning View-Invariant Features for Person Identification in Temporally Synchronized Videos Taken by Wearable Cameras. In *ICCV*.

[51] Zhun Zhong, Liang Zheng, Zhedong Zheng, Shaozi Li, and Yi Yang. 2018. Camera Style Adaptation for Person Re-identification. In *CVPR*.