International Journal of Automation and Computing
www.ijac.net

18(3), June 2021, 334-350
DOI: 10.1007/s11633-020-1258-8

A Comprehensive Review of Group Activity Recognition in Videos

Li-Fang Wu[1,2] Qi Wang[1] Meng Jian[1,2] Yu Qiao[3] Bo-Xuan Zhao[1]

[1] College of Information and Communication Engineering, Beijing University of Technology, Beijing 100124, China

[2] Beijing Municipal Key Lab of Computation Intelligence and Intelligent Systems, Beijing University of Technology, Beijing 100124, China

[3] Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China

Abstract: Human group activity recognition (GAR) has attracted significant attention from computer vision researchers due to its wide practical applications in security surveillance, social role understanding and sports video analysis. In this paper, we give a comprehensive overview of the advances in group activity recognition in videos during the past 20 years. First, we provide a summary and comparison of 11 GAR video datasets in this field. Second, we survey the group activity recognition methods, including those based on handcrafted features and those based on deep learning networks. For better understanding of the pros and cons of these methods, we compare various models from the past to the present. Finally, we outline several challenging issues and possible directions for future research. From this comprehensive literature review, readers can obtain an overview of progress in group activity recognition for future studies.

Keywords: Group activity recognition (GAR), human activity recognition, scene understanding, video analysis, computer vision.

Citation: L. F. Wu, Q. Wang, J. Meng, Q. Yu, B. X. Zhao. A comprehensive review of group activity recognition in videos. *International Journal of Automation and Computing*, vol.18, no.3, pp.334–350, 2021. http://doi.org/10.1007/s11633-020-1258-8

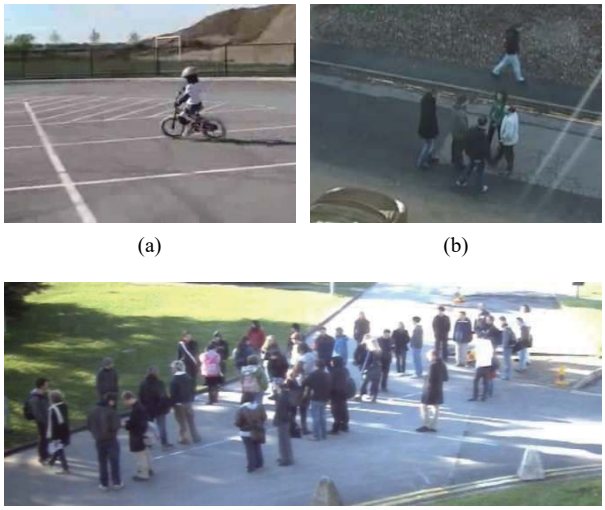
1 Introduction

In recent years, the widespread applications of surveillance equipment have rapidly increased the amount of video data. Analyzing and understanding the complicated video contents has become an urgent demand. Human activity analysis, as a challenging research topic for video contents analysis, has attracted intensive research interest in the community of computer vision. In previous decades, human activity analysis has made remarkable progress.

Human activity is a complicated concept and there are various levels. To present our work clearly, we categorize human activities into three different levels based on the complexity: individual action, group activity and crowd behavior. Fig. 1 demonstrates instances of these three levels. Individual action covers single-person action where the human pose and the motion of human body are discriminative information. The crowd behavior is occurred at environment with high dense crowds. Thus, it is infeasible to obtain the precise tracks and detailed information about an individual person. The objective of research in human crowds lies in identifying abnormal activity or emergency situations based on the motion pattern of crowds.

In this paper, we focus on group activity which is composed of one or more sub-groups involving visually countable persons with interactions in the scene. A distinctive property of group activity recognition is the interactions between different groups and individuals. As illustrated in Figs. 2(a) and 2(b), two highlighted people share a similar appearance with the same atomic action "standing", however, it is ambiguous to distinguish the group activity based on the action of only a single individual. The interaction among persons in the group should be considered to infer group activity. For instance, in Fig. 2(a), two or three people standing face to face indicate they are talking while in Fig. 2(b), many people standing and facing in the same direction reveals they might queue. Fig. 3 demonstrates that only a few key individuals play important roles in the group activity and other people might bring irrelevant information. Therefore, it is reasonable to predict group activity on the basis of contextual information in the entire image rather than isolated information from a single individual. Compared to crowd behavior, group activity enables us to capture detailed information about individuals as well as their interactions, which is more easily explained and makes sense in practice. The interactions among a group of persons occur much more often in practical scenarios and the study of group activity recognition has tremendous poten-

Review
Manuscript received May 8, 2020; accepted September 25, 2020; published online January 11, 2021
Recommended by Associate Editor De Xu
Colored figures are available in the online version at https://link.springer.com/journal/11633



(a) (b) (c)

Fig. 1 Three levels of human activity analysis: (a) Human action; (b) Group activity; (c) Crowd behavior.

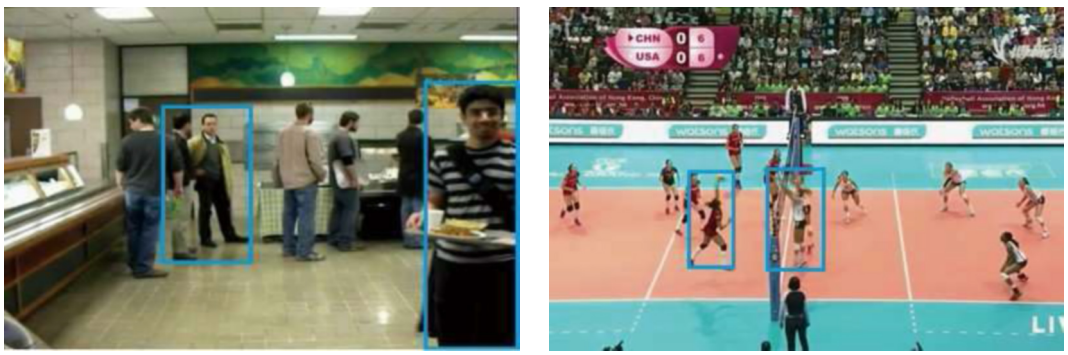
tial for many applications such as sport video analysis and smart video surveillance.

To sum up, group activity recognition is of theoretical and practical significance. However, most of previous reviews of human activity recognition are focused on individual action recognition[1–4]. Reviews about group activity recognition are scarce. Fauzi and Sulistyo[5] mainly survey the connection between group activity recognition and the advancement of internet of things (IoT) technology in smart buildings. Aggarwal and Ryoo[6] study different levels of human activity, however they only introduce traditional methods. To the best of our knowledge, the most recent survey related to group activity recognition is published in 2017[7]. It focuses mainly on handcrafted based methods while deep learning based methods are not discussed in depth. Moreover, notable progress has been made in this field in recent years because of powerful deep learning techniques. Therefore, an overview of group activity recognition methods including the state-of-the-art in recent years is required. Compared with previous surveys, our survey introduces sufficient latest works and discusses recent research trends in group activity recognition.

This paper provides a comprehensive survey of current group activity recognition methods. We distinguish between the traditional approaches based on handcrafted features and those based on deep learning. For traditional methods, we further divide them into two categories. The first is the top-down approach which relies on analyzing the group-level information to recognize activity. The second one is bottom-up approach which recognizes activity based on each individual in group contexts. For approaches based on deep learning, we categorize four classes on the basis of what crucial problem they focus on. We also give a summary of publicly available data-

(a) (b)

Fig. 2 Role of contextual information. The group activity in (a) is talking, the group activity in (b) is queuing. Two highlighted people performing different actions share similar appearance features. These two pictures demonstrate that the interaction between individuals is a crucial cue for recognizing group activity.



(a) (b)

Fig. 3 The group activity is usually determined by a few key individuals: (a) The main group activity is queuing while a person in the right of the image is walking and some people are talking in the queue. (b) The group activity is left spiking. The spiking player and blocking players are leading the group activity. These two pictures indicate that each group might perform different activities and it is essential to consider the contextual information of the whole scene to infer group activity.

sets and the comparisons between state-of-the-art approaches.

This paper is organized as follows. A dataset summary is provided in Section 2. In Section 3, traditional approaches are divided into two categories and each category is reviewed with a specific description. In Section 4, deep learning based approaches proposed in recent years are detailed introduced. Finally, Sections 5 and 6 describe the research challenges and conclusions respectively.

2 Datasets

The public datasets provide a unified measurement and direct comparison for proposed methods, which leads to better understanding of the pros and cons of each algorithm. Therefore, constructing datasets plays an essential role for promoting the development of group activity recognition. Compared to benchmarks available for understanding individual actions, there are few resources involved in complex human group activities. All the datasets for group activity recognition belong to surveillance videos or sports videos which are motivated by the practical requirements for constructing safety systems or sports analysis systems. In this section, we provide an overview of available datasets. Example video frames appear in Fig. 4. A summary of datasets appears in Table 1.

2.1 Surveillance datasets

All the surveillance datasets are collected in practical environments such as the campus or street. Most of the

Fig. 4 Example video frames from different datasets: (a) The BEHAVE Dataset[8]; (b) The Collective Activity Dataset[9]; (c) The Volleyball Dataset[10]; (d) The NCAA Basketball Dataset[11].

Table 1 Summary of datasets

Dataset	Group activity category	Atomic action category	Year	Footage type	Best accuracy achieved
NUS-HGA	6	N/A	2009	Surveillance video	91.7%Cheng et al.[12]
BEHAVE	10	N/A	2009	Surveillance video	77.6%Zhang et al.[13]
CAD1	5	6	2009	Surveillance video	95.7% Tang et al.[14]
CAD2	6	8	2011	Surveillance video	85.5% Khamis et al.[15]
CAD3	6	3	2012	Surveillance video	87.2% Amer et al.[16]
UCLA Courtyard	6	10	2012	Surveillance video	83.7% Amer et al.[17]
Nursing Home	2	6	2012	Surveillance video	85.5% Deng et al.[18]
Broadcast Field Hockey	3	11	2012	Sports video	62.9% Lan et al.[19]
NCAA Basketball	11	N/A	2016	Sports video	58.1% Wu et al.[20]
Volleyball	8	8	2016	Sports video	94.4% Gavrilyuk et al.[21]
C-Sports	5	N/A	2020	Sports video	81.3% Zalluhoglu and Ikizler-Cinbis[22]
NBA	9	N/A	2020	Sports video	47.5% Yan et al.[23]

videos are recorded with a stationary monitor indicating that the backgrounds are static without camera motion. Background clutter and occlusions between multiple people occur frequently.

NUS-HGA Dataset[24] is collected by a monitor at university car park. This dataset consists of six different group activities: Walk in Group, Ignore, Gather, Stand and Talk, Fight and Run in Group. Each activity clip takes 8–15 seconds with 4–8 actors. The dataset has 476 labeled video samples in total.

BEHAVE Dataset[8] consists of 10 types of group activity classes: InGroup, Approach, WalkTogether, Meet, Split, Ignore, Chase, Fight, RunTogether and Following. There are usually 2–5 people as a group or two groups interacting in each video. This dataset contains 174 samples of different group activities and in total 76 800 individual frames.

Collective Activity Dataset(CAD1)[9] is one of the widely used benchmarks for group activity recognition. It contains 44 short video clips from 5 activity categories (Crossing, Waiting, Queueing, Walking and Talking) and 6 individual action categories (NA, Crossing, Waiting, Queueing, Walking and Talking). The group activity is labeled for a clip by the activity in which most people participate. The benchmark also provides 8 pose orientation labels, 8 pairwise interaction labels and trajectory of each person in video clip. The above annotations are manually labeled every ten frames.

Collective Activity Extended Dataset (CAD2)[25] augments the Collective Activity Dataset[9] by adding two more categories of dancing and jogging as a new class and removing the Walking activity as the Walking activity is an individual action rather than a group activity. The Collective Activity Extended Dataset contains in total 75 video sequences.

New Collective Activity Dataset (CAD3)[26] is comprised of 32 video clips with 6 group activities: gathering, talking, dismissal, walking together, chasing and queueing. Three atomic actions are labeled as walking, standing still and running, and 9 interaction labels are defined.

UCLA Courtyard Dataset[27] contains 106-minute high-resolution videos from a bird-eye viewpoint of a courtyard at the UCLA campus. The annotation of datasets provides 6 group activities (Walking-together, Standing-in-line, Discussing-in-group, Sitting-together, Waiting-in-group and Guided-tour) and 10 individual actions.

Nursing Home Dataset[28] consists of videos captured in a dining room of a nursing home by fixed low-resolution surveillance camera. Individual actions include walking, standing, sitting, bending and falling. Based on the individual actions, each frame is assigned by two activity categories: fall and non-fall. If any person falls, the frame is assigned "fall", vice versa. In total, there are 22 short video clips and 2 990 annotated frames in this dataset.

2.2 Sports datasets

Sports datasets are usually collected from broadcast video. In most of the cases, the camera moves with the occurrence of some specific event. Compared with surveillance datasets, sports datasets have more complicated person-person interactions and heavy occlusions. Moreover, the sport activities are usually sensitive to a few players such as the spike event in volleyball game is mainly determined by the spiking player and blocking players.

Broadcast Field Hockey Dataset[19] has 58 video sequences with 11 atomic actions: pass, dribble, shot, receive, tackle, prepare, stand, jog, run, walk and save, and 3 scene-level events: attack play, free hit and penalty corner. Besides, to explore the effect of social roles on group activity, five social roles are defined.

NCAA Basketball Dataset[11] collects 257 NCAA basketball games available from YouTube and each untrimmed video is 1.5 hours long. Eleven key events are defined including 5 types of shots, each of which could be successful or failed, plus additional a steal event. This dataset is challenging due to heavily mutual occlusion, low resolution and the complicated interactions in sports video.

Volleyball Dataset[10] is a more challenging dataset due to large scale, complicated interactions and rapid motion of players. This dataset is collected from available volleyball game videos in YouTube. It consists of 4 830 video clips gathered from 55 games. Each clip is only annotated in the middle frame in which each player is labeled by a bounding box with individual actions and a group activity category is provided for each clip. There are a total of 8 group activity categories (Left/Right set, Left/Right spike, Left/Right pass and Left/Right winpoint) and 8 individual atomic actions (Waiting, Setting, Digging, Falling, Spiking, Blocking, Jumping, Moving and Standing).

C-Sports[22] is a benchmark for multi-task recognition of both group activity and sports categories. In this dataset, there are 11 types of sports and 5 group activity categories. Sports categories include American football, basketball, dodgeball, football, handball, hurling, ice hockey, lacrosse, rugby, volleyball and water polo. Group activities are gathering, dismissal, passing, attack and wandering. To estimate the generalization ability of the algorithm, a challenging evaluation protocol in C-Sports is introduced which training and testing are on different sport classes respectively.

NBA Dataset[23] is currently the largest and the most challenging benchmark for group activity analysis. Unlike conventional GAR tasks, this dataset presents a new task namely weakly-supervised group activity recognition in which person-level information is not provided even in the training data and only video-level labels are available. It collects 181 NBA games from the web and

there are 9 172 video clips, each of which belongs to one of the 9 activities.

3 Approaches based on handcrafted features

Traditional approaches to group activity recognition can be categorized into two classes: top-down approaches and bottom-up approaches. The top-down approaches analyze activities in terms of group level motion and interaction. The drawbacks for these approaches are a lack of detail description for activity that they cannot fully exploit features at individual level. The bottom-up approaches focus on recognizing each individual and describing the activity based on a collection of individual features and their statistics. Therefore, they are sensitive to individual feature extraction failure due to occlusion or missed detection. This section reviews both types of approaches and we compare top-down and bottom-up approaches in Tables 2 and 3 respectively.

3.1 Top-down approach

Top-down approaches are focused on analysis of global motion patterns of an entire group or each sub-group and investigate the trajectory as well as interaction of groups while the individual action of a specific actor in the scene is less important. In this way, they are more robust to occlusion and low-resolution.

Table 2 Comparisons of top-down approaches

Approach	NUS-HGA	BEHAVE	Others
Ni et al.[24]	73.5	70.64	
Blunsden and Fisher[8]		74.66	
Cheng et al.[12]	91.7		
Zhang et al.[13]		77.6	
Tran et al.[29]	79.1		CAD1: 78.75
			CAD2: 80.77
Cheng et al.[30]	96.2		

Table 3 Comparisons of bottom-up approaches

Approach	CAD1	CAD2	CAD3	Others
Choi et al.[9]	65.9			
Choi et al.[25]	70.9	82.0		
Choi and Savarese[26]	79.0		83.0	
Amer et al.[27]	83.6			UCLA: 72.7
Lan et al.[31]	68.2			
Kaneko et al.[32]	73.2			
Nabi et al.[33]		72.9	72.3	
Lan et al.[34]	77.5			
Chang et al.[35]	83.3		80.3	
Amer et al.[17]	88.9		84.2	UCLA: 83.7
Amer et al.[16]	92.0		87.2	
Hossein et al.[36]	83.4			
Khamis et al.[15]	72.0	85.8		

3.1.1 Trajectory based method

Trajectory based methods are centered on analyzing group activities in terms of interactions between individual trajectories. Vaswani et al.[37] model moving objects as point objects in the two-dimension plane. Instead of tracking each point and recognizing their interaction, they propose to represent a group activity as the polygonal shape change of these points′ configuration over time frames following the Kendall′s shape theory. At each time, they extract object points in the image and construct a polygon based on these points. A tangent coordinate system is defined as the mean shape which is learned from observed object configurations from a training sequence of frames. The normal and abnormal activities are distinguished by comparing the extracted shape from input frames with the learned model in the tangent space. Similarly, Khan and Shah[38] proposed a method to detect group activities which can be characterized by rigidity information such as parading or marching. They represent each entity as a corner of three-dimension polygons and the tracklets of each entity on the three-dimension polygon plane are treated as trajectory feature. The final classification results are inferred from the structure composed from the trajectory and interactions between participating entities.

Zhou et al.[39] designed a set of features which measures the strength of causality between two trajectories and another set describes the type of causality. The two sets of features along with conventional velocity and position features of a trajectory-pair are fused to explore the relationship between two object entities. However, they can only deal with the pair-activity recognition. Ni et al.[24] proposed to analyze group activity with self-causality, pair-causality and group-causality based on local trajectory information. These three categories of causality extract dynamic interaction relations of different individuals and describe the spatial and temporal characters of behaviors of the human group. Cheng et al.[12] introduced Gaussian processes to describe motion trajectories of individuals and provide a probabilistic perspective on explaining the variation of individual in group. Three descriptors, namely Individual, Dual and Unitized Group Activity Pattern respectively, are designed to capture relationships of individuals in group activities. Zhang et al.[40] proposed to obtain group-level context from extracted individual trajectories. They constructed a weighted graph to represent the probabilistic group membership of the individuals. The features extracted from this graph can capture the motion and action context for group

event recognition.

3.1.2 Sub-group interaction

To cope with complicated situations where multiple groups perform different activities in a scene, some methods detect sub-groups firstly then analyze the interactions of different groups and the activity of each group. Yin et al.[41] first clustered each individual into several sub-groups by the minimum spanning tree algorithm and then used social network analysis based feature description to extract structural features which contain the global pattern of each sub-groups as well as local motion information of the individual in each group. Finally, a Gaussian process dynamical model is trained to model different group behaviors respectively. Zhang et al.[13] proposed to represent group behavior with a combination of subgroups and introduced multi-group causalities: individual, pair, behavior and inter-group causality to describe the interaction between groups. Furthermore, they employed an improved locality-constrained linear coding method to encode the proposed multi-group causalities. Azorin-Lopez et al.[42] proposed a descriptor vector which describes not only the trajectory of individuals in a group, but also the trajectory followed by sub-groups and the movement relationship between different sub-groups in the scene. The trajectory analysis provides a path to understand complex high-level groups activities.

Sub-group information is a helpful cue for recognizing group activity under complicated scenes, however how to identify meaningful sub-group remains a challenging problem. Kim et al.[43] proposed to detect the group interaction zone and update it over time so that noisy information can be suppressed and the active zone for activity can be enhanced. To represent interactions within group interaction zones, they further proposed two features, group interaction energy feature, attraction and repulsion features. Tran et al.[29] measured degrees of interactions between individuals by social signal cues. Then they leveraged graph clustering algorithm to discover interacting sub-groups in the scene and discarded non-dominant groups. To better understand group activity, they proposed a descriptor which encodes social interaction cues and motion information of individuals within the active sub-groups. Sun et al.[44] proposed a latent graph model to solve two tasks: group discovery and activity recognition simultaneously. They constructed a relation graph which encodes the context relations between tracklets, intra-group interaction and inter-group interaction. The model can propagate message between various layers of the latent graph.

3.1.3 Multi-camera context

Nowadays, multi-camera surveillance systems which provide larger view are set up in almost public places such as campus and airport. Therefore, there is high demand for addressing group activity recognition under multiple cameras scenarios and some researchers studied this topic. In [45], multiple tracks in multi-cameras are used to extract spatio-temporal features of individuals. They considered two hierarchical clustering approaches for grouping individuals, agglomerative clustering and decisive clustering, using dissimilarity to measure between tracked targets. Zha et al.[46] proposed a graphical model with hidden variables from which intra-camera and inter-camera contexts are extracted. By optimizing the structure of graphical model, the contexts are explored automatically. Moreover, they present a spatio-temporal feature, namely vigilant area, to encode the motion information in an area which is proven to be effective for group activity representation.

3.1.4 Discussions

Trajectory-based methods are based on the observation that the tracking of individual positions and the overall movement of group are sufficient for recognizing group activity. Therefore, trajectory-based methods are suitable for recognizing the group activity which is characterized by the overall motion of an entire group. However, most trajectory-based methods focus on the activity with only one group without considering the fact that the group behavior in the real scenario usually consists of multiple groups and is mainly characterized by the dynamic interaction among groups of individuals. Sub-group interaction based methods address this problem by detecting sub-groups and utilizing interaction information among groups to better understand group activity. The major advantage of such approaches is their ability to analyze the interactions of groups. Unlike the aforementioned methods, multi-camera context based methods predict group activity with multiple cameras. In multi-camera scenes, intra-camera and inter-camera contexts are important information. In general, top-down approaches are analyzing activities in terms of group level motion and interaction and not heavily relying on individual feature which are robust to occlusions or low-resolution. Comparison between top-down approaches is shown in Table 2.

3.2 Bottom-up approach

Bottom-up approaches can be applied for recognizing group activity with a limited number of people in the scene who have nonuniform behaviors. For example, in indoor environments such as coffee shop, some people are talking face to face while other people are queuing for ordering coffee or just standing. These types of approaches usually recognize each individual person and then analyze their hierarchical structure: individual level and group level.

3.2.1 HMM based model

In the previous studies, hidden Markov model (HMM) is applied to address structure data in video. Zhang et al.[47] recognize group activity for meetings including monologues, discussion, presentation and note-taking. They proposed a two-layer hidden Markov model in

which the first layer models basic individual action by low-level audio-visual features, and the second layer models the interaction between meeting participants. Similarly, Dai et al.[48] recognize break, presentation and discussion in meeting scenarios using event-driven multi-level deep belief nets (EDM-DBN) which models group interactions as a group of Markov chains.

3.2.2 Descriptor based method

Later some researchers combined context information by designing new descriptors extracted from individual or surrounding scenes to model the evolution of group activity. Choi et al.[9] introduced a spatio-temporal local (STL) descriptor which calculates the spatial temporal distribution of position, pose and motion information of individuals. The STL descriptor is centered on an anchor person and captures histograms of surrounding persons with their poses and motion information in different bins. Choi et al.[25] extends the STL descriptor and proposed randomized spatio-temporal volume (RSTV) representation. The framework is built upon a random forest structure which randomly samples portions of spatio-temporal volume and the discriminative regions for classification. This method can automatically discover the optimal configuration of spatio-temporal bins so as to increase discriminating ability of the algorithm.

Motivated by the fact that what other surrounding people doing is a constructive cue for analyzing the actions of each individual. Lan et al.[31] proposed the action context (AC) descriptor which captures the actions of the anchor person as well as other people nearby. Experimental results demonstrate that this method can deal with complex activities in a surveillance scene. However, the AC descriptor is sensitive to viewpoint change. To solve this problem, Kaneko et al.[32] proposed the relative action context (RAC) descriptor which encodes relative relation and is invariant under viewpoint change.

To make the low-level feature extractors provide more discriminative information for high-level inference models, Amer and Todorovic[49] introduced a mid-level feature descriptor bags-of-right-detections (BORD) which seeks to discover individuals who participate in group activity and remove irrelevant individuals in groups. Specifically, the BORD descriptor is a histogram of human poses which calculates with people who participate in the activity. The chains of BORDs are fed into a two-step maximum a posterior (MAP) inference to construct activity representation.

Existing methods heavily depend on the accuracy of detectors that might fail in the crowd scenarios due to occlusion. Nabi et al.[33] presented a semantic-based spatio-temporal descriptor based on Poselet activation patterns over time. This descriptor is designed for modeling human motion interactions in crowded cases. Experiential results revealed that this descriptor can effectively tackle complex real scenarios in group activity recognition and activity localization.

3.2.3 Interaction context

In addition, one of the essential properties of group activity is relationship and interactions between individuals, including person-person interactions, person-group interactions and group-group interactions, which are useful cues to reason about group activity. Lan et al.[34] introduced a hierarchical interaction model and an adaptive interaction structure mechanism to automatically search for the suitable structure to infer activity. Finally, the person-person interaction only builds between the subset of relevant people. Kaneko et al.[50] proposed to utilize fully connected CRFs to integrate multiple types of individual features such as position, size and motion. Thus, different shapes and types of groups can be handled. Chang et al.[35] focused on modeling the person-person interaction. They utilized the features of individuals in pairs and modeled relations peer-to-peer. The interaction pattern is obtained via the interaction matrix which is learned by maximizing the interaction responses.

Graphical models and their variants are commonly used tools for group activity recognition. Amer et al.[27] proposed a graph based interaction method. An AND-OR graph is present to model objects occurring in the scene, individual action and group activity simultaneously. They proposed a principled formulation for efficient graph inference by an explore-exploit strategy. In [17], they further proposed a hierarchical, spatio-temporal AND-OR graph (ST-AOG) which models both individual actions, group activities and relations of individual actions within a group activity. Moreover, Monte Carlo tree search is used to address expensive computation cost on AOG inference. Later, Amer et al.[16] advanced the existing graph model with a hierarchical random field (HiRF). HiRF is designed for extracting spatio-temporal features in video and capturing long-range dependencies. HiRF aggregates multi-scale input features and discovers foreground features of groups, while removes features that belong to background clutter.

Lan et al.[19] utilized social roles to complement the representation of low-level individual and high-level events within a graph framework. In the proposed graphical model, individual action is modeled based on individual feature vectors at the lowest level and the contextual interaction information between individuals are modeled based on their social roles at the intermediate level. Group-level events are inferred at the top level of model. Zhao et al.[51] observed that most existing approaches assumed all individuals share the same activity label and ignore multiple activities co-existing in some scenarios. This factor can serve as a context cue in many cases. They present a unified discriminative learning framework of multiple context models which takes both the intra-class and inter-class behavior interactions among persons into consideration. Activities always have serious intra-class variation caused by changes of individual appearance or temporal evolution, which will lead

to confusion for the recognition algorithms. To solve this, Lan et al.[52] presented a method which additionally models action primitives and considers the interactions of theirs. Sometimes, activity of a group of people can be classified by counting the actions of individual in the scene. Hajimirsadeghi et al.[36] developed a probabilistic structured kernel method that is based on the multi-instance model to infer cardinality relations which can reduce the confusion caused by irrelevant individuals. The results show that encoding cardinality relations can obtain significant improvements on performance for group activity classification. Zhou et al.[53] addressed the problem of recognizing mixed group activities contained in one still image. They proposed a four-level structure which captures interactions among group to group and interactions among person to person. Experimental results demonstrate that the model is robust to scenes with high crowd density and can well tackle the problem of the mixed group activities.

3.2.4 Tracklets based method

For bottom-up approaches, an integral step is identifying coherent trajectories of each individual. However, tracking multiple individuals at the same time is challenging because of self-occlusion, background clutter or camera shaking. These factors lead to inaccurate tracklets which are not stable enough to construct the recognition algorithm. Most approaches isolate tasks of the tracking and recognizing activity. Choi and Savarese[26] presented a framework for simultaneously tracking multiple individual and estimating group activity. What underlies the intuition of treating the two problems jointly is that persons′ motion and their activity have a strong correlation. Performing the two tasks in a coherent fashion means that the two components can promote each other. They exploited interactions between individuals for guiding the target associating process and designed a hierarchical graphical model to encode the correlation between activities. Khamis et al.[15] is motivated by the discordance of an action in a scene. Sometimes, an object performing different actions may share similar appearance for frame-level features and different motion information in the track level feature. They proposed a model which captures the relevance between individual′s action and the motion flow in the video sequence. Finally, group activities are inferred by combining per-frame and per-track cues.

3.2.5 Discussions

Bottom-up approaches are suitable for recognizing the group activity with a limited number of members who have their own role, different from the others. For example, the group activity, presentation in a meeting room: the presenter is talking while the other members are listening or taking notes. This type of group activity requires methods to have the ability of recognizing actions of each individual and their structures. The HMM-based model is applicable to address hierarchical structure. At the bottom layer, atomic actions of individuals are recognized from sequences while the second-level layer models activities of the group. Context information in the scene is helpful to differentiate ambiguous activities such as standing and queuing. Descriptor based methods propose various feature descriptors extracted from a focal individual and its surrounding area to integrate contextual information. Unlike the descriptor based methods which provide context information between focal individual feature with all people within group, the interaction context model provides interaction information among person to person, person to group and group to group which makes it possible to tackle complicated interaction scenarios. For bottom-up approaches, identifying coherent trajectories of each individual is a preprocess step for group activity recognition. Previous methods are isolating tasks of the tracking and recognizing, however a person′s motion and their activity are sometimes correlated. The goal of the tracklets based method is performing two tasks jointly and making them promote each other. A comparison between bottom-up approaches is shown in Table 3.

4 Deep learning based methods

Recently, deep convolutional neural networks (CNNs) have demonstrated impressive performance on a variety of computer vision tasks including image classification[54], semantic segmentation[55], image super-resolution[56] and video recognition[57]. Several deep learning approaches have been proposed for group activity recognition and achieved superior results to handcrafted approaches. This section reviews deep learning based methods for group activity recognition. We summarize four key problems for group activity recognition: hierarchical temporal modeling, relationship modeling, attention modeling and a unified modeling framework. We divide methods based on what crucial problem they focus on. The comparison results for deep learning based methods are demonstrated in Table 4.

4.1 Hierarchical temporal modeling

The group activity recognition needs to simultaneously reason on a collective of persons. A challenge for this task is how to design appropriate networks to allow the learning algorithm to focus on differentiating higher-level classes of activities which are about spatial and temporal evolution of the group activity. Long short-term memory network (LSTM)[70], a particular type of recurrent neural network, has achieved great success in sequential tasks including speech recognition[71] and image captioning generation[72]. For group activity recognition, some researchers attempt to apply LSTM to construct a hierarchical structure representation to infer individual actions and group activities[10, 20, 58–60, 73–76].

Ibrahim et al.[10] proposed a two-stage hierarchical

Table 4 Comparisons of deep learning based approaches. For CAD1, multi-class accuracy (MCA) and mean per class accuracy (MPCA) are shown

Dataset	Category	CAD1(MCA/MPCA)	Volleyball dataset	Others
Ibrahim et al.[10]	Hierarchical temporal modeling	81.5/89.7	81.9	
Wang et al.[58]	Hierarchical temporal modeling	N/89.4		CAD3: 85.2/89.4
Shu et al.[59]	Hierarchical temporal modeling	87.2/88.3	83.3	
Li and Chuah[60]	Hierarchical temporal modeling	86.1/89.0	66.9	
Deng et al.[18]	Deep relationship modeling	81.2/N		CAD2: 90.23/N Nursing home:85.50
Qi et al.[61]	Deep relationship modeling	89.1/N	89.3	
Ibrahim and Mori[62]	Deep relationship modeling		89.5	
Azar et al.[63]	Deep relationship modeling	85.75/94.2	93.04	
Wu et al.[64]	Deep relationship modeling	91.0/N	92.6	
Hu et al.[65]	Deep relationship modeling	N/93.8	91.4	
Gavrilyuk et al.[21]	Deep relationship modeling	92.8/N	94.4	
Yan et al.[66]	Attention modeling	N/ 92.2	87.7	
Tang et al.[14]	Attention modeling	N/95.7	90.7	
Lu et al.[67]	Attention modeling	90.6/N	91.9	CAD2: 91.2/N CAD3: 89.2/N
Bagautdinov et al.[68]	Unified modeling framework		87.1	
Zhang et al.[69]	Unified modeling framework	83.8/N	86.0	

deep temporal model (HDTM). The first stage applies a person-level LSTM to the tracklets of each individual to model individual activities. In the second stage, a group-level LSTM is adopted to combine individual-level information and form group level features for group activities. This method is the first work that incorporates a deep LSTM framework to address group activity recognition.

Besides person-person and person-group interactions, group activity is often associated with interactions between sub-groups. Wang et al.[58] proposed a multi-level interaction context encoding network on the basis of a hierarchical LSTM framework[10]. The network models three level interactions including individual dynamics, intra-group individual interactions and inter-group interactions. To enrich person level features, they deployed a contextual binary encoder which encodes the sub-action in the framework.

Shu et al.[59] argued that existing group activity recognition benchmark datasets (the collective activity dataset[9] and the volleyball dataset[10]) are too small to train a robust LSTMs framework. To solve this problem, they proposed the confidence-energy recurrent network (CERN) which extends the two-level hierarchy of LSTMs framework by incorporating a confidence measure and an energy-based model.

Inspired by the fact that people can infer an activity from a sequence of sentences easily, Li and Chuah[60] presented a semantics-based scheme, namely SBGAR. They designed a LSTM model to generate a caption for each video frame in the first stage. In the second stage, another LSTM model predicts group activities based on the generated caption of a sequence of frames. This is the first cross-modal method for group activity recognition and achieved the state-of-the-art results at that time.

Sometimes, different group activities share the same local motion which may cause misclassifications. To reduce the influence of confused motions, Kim et al.[73] proposed a discriminative group context feature (DGCF) that takes prominent sub-events into consideration. Two types of features, individual activity and sub-event feature, are extracted to construct group activity representations. The model is based on the gated recurrent units (GRU) model, which is a modified model of LSTM, to capture the temporal evolution in a video.

Gammulle et al.[74] presented a multi-level sequential generative adversarial network (MLS-GAN) based on LSTM architecture. This method is the first attempt to introduce GAN to the group activity recognition task. Instead of utilizing manually annotated individual actions, this approach automatically learns appropriate sub-actions which are pertinent to the final group activity by generative adversarial networks, within which the generator, trained with sequences of person-level and scene-level features, learns an action representation and the discriminator performs group activity classification.

Wu et al.[75] proposed global motion pattern to represent complex multi-person motions in the sports video.

Global motion patterns extracted by an optical flow algorithm are fed into convolutional neural networks and LSTM networks to extract spatial and temporal features for event classification. They further extend the GMP framework in [20]. A two-stage scheme for event classification in basketball videos is proposed. In the first stage, event occurrence segments and post-event segments are utilized for event classification and the failure/success of an offense respectively. Eventually, final results are obtained by the integration of event classification results and success/failure classification results.

Previous two-stage LSTM based methods neglect the fact that person-level actions and group-level activity are occurring over time. To this end, Shu et al.[76] proposed a graph LSTM-in-LSTM (GLIL) network which jointly models the person-level actions and the group-level activity. Multiple P-LSTMs model the person-level actions based on the interactions among individuals. Meanwhile, a G-LSTM models the group-level activity and the person-level information in P-LSTMs is selectively integrated into G-LSTM.

4.2 Deep relationship modeling

Building relationships between persons and performing relational reasoning are essential for recognition of higher-level activities. However, modeling relevant relations between people is challenging in group activity recognition for the reason that only individual action labels and group activity labels are accessible, without additional knowledge of interaction information. Much research[18, 21, 61–65, 77–83] explores how to capture the contextual information about the person in the scene and their relations.

Deng et al.[77] focused on modeling the interaction between individuals and their relationship in the scene. This is achieved by a multi-layer perceptron for capturing the dependencies of individual actions, group activity and scene labels. They further proposed a structure inference machine[18] which is consisted of a deep convolution network with a graphical model. They utilized a recurrent neural network to propagate messages between individual people in a scene. Moreover, a trainable gating function is designed to suppress the influence of irrelevant people in the scene.

Qi et al.[61] proposed an attentive semantic recurrent neural network, namely stagNet. A semantic graph is built from word labels and visual data. Individual actions and temporal contextual information are integrated by a structural-RNN model. The spatial relationship between individual people is inferred in a semantic graph via a message passing mechanism. Beyond that, person-level spatial attention and frame-level temporal attention are designed to automatically discover the key person and the key frame.

To acquire a compact relational representation of each individual person, Ibrahim and Mori[62] developed the relational layer that refines relationship representations based on a relation graph. In the relational layer, each pair of individual features is aggregated by a shared neural network into a new relation to represent their relationship. By stacking multiple relational layers, a compact group representation encoding hierarchical relationships of interaction is obtained.

Existing methods have not thoroughly explored the spatial relationship between persons. To address this issue, Azar et al.[63] proposed a novel spatial representation, dubbed an activity map, based on individual and group activities. Motivated by [84], the activity map is refined in multiple stages for decreasing the incorrect representations. An aggregation method ensures the refined activity map can produce reliable group activity labels.

Graph convolutional networks (GCN)[85] have become an emerging topic in deep learning. GCNs have been applied to many fields of computer vision such as visual tracking[86] and single human action recognition[87, 88]. Graph convolutional networks are suitable model to address group activity recognition within which each person can be regarded as a node. Wu et al.[64] introduced GCN into group activity recognition. Person-level features are extracted by convolution neural networks and an actor relation graph are built based on visual similarity and spatial location distance between individual persons. Graph convolution networks are adopted to perform relational reasoning on the actor relation graph to acquire the relational features of each person.

Hu et al.[65] applied deep reinforcement learning for relation learning in group activity recognition which is a new method. A semantic relation graph is built to model relations of persons in the scene. Then, two agents based on Markov decision processes are applied to refine the graph. The relation gating agent is responsible for enforcing relevant relation learning and discarding irrelevant relations. Another feature-distilling agent distills the key frames of features which is similar to a temporal attention mechanism.

Xu et al.[79] proposed a multi-modal relation representation with temporal-spatial attention which infers relations from appearance features and motion information. Two types of inference modules, opt-GRU and relation-GRU, which are used to encode the object relationship and motion representation effectively, are introduced to form the discriminative frame-level feature representation.

Inspired by a transformer network[80] which relies on self-attention mechanisms to allow the network to adaptively extract the most relevant information and relationships, Gavrilyuk et al.[21] proposed an actor-transformers network which learns interactions between the actors and adaptively extracts the important information for activity recognition.

In the real scene, individuals may perform their own actions or they might be connected to a social group and several groups of people have potentially different social

connections. Ehsanpour et al.[81] proposed a new task social activity recognition which simultaneously performs individual action prediction, social group division and sub-group activities predicting.

4.3 Attention modeling

For group activity recognition, there are usually several persons active in the scene while only a few key persons are contributing to group activities, and others who may bring confusing information for inferring group activities are irrelevant. Due to lack of key person annotations for group activity recognition datasets, this problem can be defined as weakly supervised important people detection. To address this issue, several methods[11, 14, 66, 67, 89–92] designed attention mechanism.

Ramanathan et al.[11] worked on basketball event detection which is sensitive to a subset of players. They formulated a spatial and temporal attention model to attend relevant players for events in the scene and apply weighted summation mechanisms to extract person-level features which lead to a better representation for event detection.

Yan et al.[66] observed that the actors who move steadily during the whole process or move remarkably at a moment have more contributing to the group activity. To measure the mean motion intensity which represents long motion of an actor, they stack the optical flow images of the video clip and calculate the mean intensity of them. The intensity of flash motion for an actor is captured by learning an attention factor to weight sum of his/her hidden state from LSTM at every time step. In [86], they further proposed a coherence constrained graph LSTM with a temporal confidence gate and a spatial confidence gate to control the memory updating. Meanwhile, an attention mechanism is constructed to measure the contribution of a motion at each time step.

Previous methods address key actor detection by self-attention mechanisms which are unreliable and lack interpretability. Tang et al.[14] provided a new insight on designing attention networks for group activity. A teacher network in the semantic domain is designed to recognize group activities based on the words of individual action labels. Then they train a student network in the visual domain to infer group activities based on video clips. In the training process, the teacher network distills attention knowledge into a student network, which is effective to mine the key people and suppress the irrelevant people without requirements for extra labels. In [90], they extended the teacher network and the student network with two types of graph neural network. By the graph convolutional modules passing the messages of different nodes, the relationship among different people in the scene can be explored.

Lu et al.[91] proposed a two-level attention mechanism for group activity recognition. The first individual-level attention is guided by pose features to control the hidden state at each time step. The second scene-level attention attaches individuals with different weights to construct discriminative scene representation. This method depends on pose estimation. Lu et al.[67] improved it and proposed a graph attention interaction model with graph attention blocks to capture unbalanced interaction relations at the individual and group level.

4.4 Unified modeling framework

Group activity recognition for video usually involves multi-person detection, multi-person tracking and activity recognition. Most existing methods separate the modeling of human detection/tracking and group activity recognition. They usually adopt an off-the-shelf human detection and tracking algorithm to preprocess the input video sequences. Their focus lies in designing a high-performance structure model to classify activity recognition. However, such practice has several drawbacks. First of all, decoupling the modeling of human detectors and group activity classifiers which ignore the inner correlation between two modules leads to suboptimal results for both parts. Second, the feature extracted by detectors for individual people is also useful for inferring group behaviors while separate learning needs extract features through backbone networks respectively which leads to extra computations.

Bagautdinov et al.[68] presented a unified framework to solve the aforementioned issues. They utilized the multi-scale feature maps output by a fully convolutional network to address three tasks: multi-people detection, individual action recognition and group activity recognition. A matching mechanism is designed for associating the same person in consecutive frames and features are fused by standard GRU in the temporal domain.

Zhang et al.[69] focused on speeding up the inference time for group activity recognition. They proposed to perform human detection and activity reasoning simultaneously in a end-to-end framework, within which a shared backbone network is exploited to extract feature. Experiments demonstrate that people who are outliers for activity can be filtered out effectively and two tasks: human detection and group activity recognition can reinforce each other. On top of that, they proposed a latent embedding scheme for building the relation of person-person and person-group interactions.

Zhuang et al.[93] explored a new representation for group activity recognition to avoid a heavy dependency on the accuracy of human detection and tracking. They proposed a differential recurrent convolutional neural network (DRCNN) which is unnecessary to take each person′s bounding-box as input and without complicated preprocess steps. Unlike existing methods where feature extraction and parameter learning are separate, DRCNN jointly optimizes the unified deep learning framework.

4.5 Discussions

Recent deep learning based methods for group activity recognition demonstrate promising improvements in performance on traditional methods. Compared with learned features, handcrafted descriptors are often not learned and quantified automatically for discrimination and their discrimination powers are usually not guaranteed. Hierarchical temporal modeling based methods use a two-stage LSTM model to learn a temporal representation of individual-level actions and apply pooling functions to individual features to generate a group-level representation. This two-stage LSTM framework inspired a lot of follow-up work. Its limitation is treating all individuals with equal importance. However, the group activity is usually defined by a few key persons in some scenarios such as in sports videos. Attention modeling based methods attempt to solve this issue and many modifications have been proposed. From Table 4, we can see that this kind of method has higher performance than hierarchical temporal modeling based methods. However, due to lack of annotation of key persons, how to learn a stable model which can accurately find key individual is still a difficult problem. Currently, relationships among entities have been widely leveraged in various computer vision tasks. Various methods of relation reasoning are introduced into group activity recognition, such as GCN and transformers. The advantage of deep relationship modeling based methods is they can capture potential interactions and relationships between persons that can effectively discriminate person and group activity. This category of methods achieves the best results in the CAD and volleyball dataset. Unified modeling framework methods attempt to perform person detection and group activity recognition jointly in a single neural network which can speed up the algorithm and bring it closer to practical applications. However, they cannot achieve the state-of-the-art recognition accuracy. Most of the existing methods directly adopt bounding box from annotations which are inaccessible in practical applications. Research on this topic is limited. Weakly-supervised group activity recognition tasks where only video-level group activity is accessible could be another direction for group activity recognition.

5 Challenges and trends

Fig. 5 demonstrates the timeline of the development in group activity recognition over the past twenty years. Despite the great success of deep learning for group activity recognition, there are still some open research issues in this field as discussed below:

1) Reliable relation representation. Relation representation matters for group activity recognition. Group activity recognition involves multiple people performing different actions and having varied interactions in a scene. Therefore, inferring group activity requires contextual reasoning about the appearance and relations of people rather than simply a combination of individual action. Under some circumstances such as sport videos, the contribution of actors are unbalanced for group activity which causes relation representation more difficult. There are some attempts to adopt self-attention mechanisms or graph neural networks for relation modeling in group activity recognition. However, previous works rely on explicitly spatial priors to build model and are limited on temporal relations. Reliable and efficient relation representation in spatial and temporal domains among actors still need to be further explored.

2) Powerful spatio-temporal representation. While 2D CNN have achieved enormous success in image recognition, they are suboptimal for video related tasks, because video is naturally a 3D spatio-temporal signal and temporal information is vital in videos. Most of the existing works for group activity recognition usually applies 2D CNN on a single frame to extract person-level features and model temporal information by recurrent neural networks on dense frames to extract group-level feature. It is worthwhile to investigate whether spatio-temporal representations extracted from 3D CNN can be beneficial for group activity representation. Optical flow, a motion information representation, can complement appearance information for CNN-based methods in individual action recognition, while they are seldom utilized in group activity recognition methods. Introducing efficient motion-related information in group activity representation should be investigated.

3) Robust human detection, tracking and recognition. Accurate detection and tracking results are

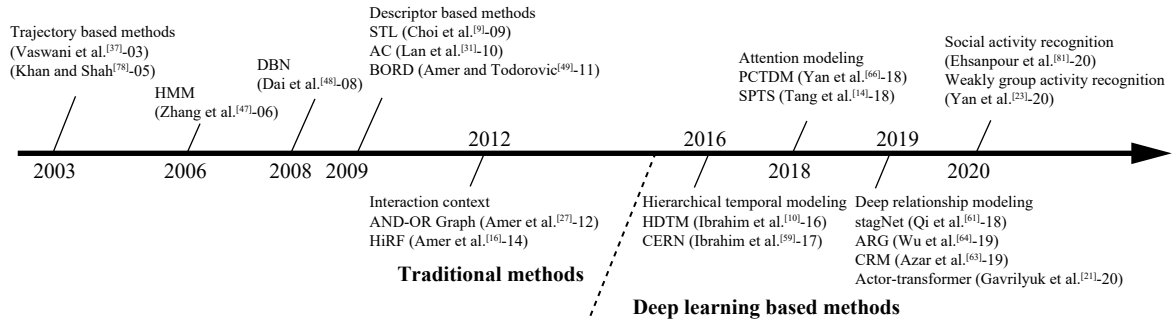


Fig. 5 Timeline of the development in group activity recognition

the fundamentals for feature extraction in high-level group activity recognition tasks. Although general detection and tracking are well-studied fields, it is challenging to detect and track multiple individuals accurately because of the frequently occurring inter-object occlusions, target-similar distractors, etc. Most of the existing methods focus on designing a structure model to classify group activity. They directly adopt a bounding box from annotations which is inaccessible in practical applications or from results of the third-party algorithm trained for general detection or tracking purpose which is non-optimal for handling multiple objects. Another attempt is to integrate human detection and group activity recognition in a unified framework which speeds up the algorithm by performing multiple tasks in a one pass-feed forward through a neural network. There are some methods working on that, but they cannot achieve the state-of-the-art classification accuracy. How to better integrate mid-level detection tasks and high-level recognition tasks is another direction of future research to further explore.

4) Bigger and challenging dataset. A brief comparison of existing collective activity recognition datasets is presented in Table 1. As it can be seen, most of the datasets are proposed before the deep learning era and these are quite limited to support the training of complex and representative models based on deep learning. The most commonly used volleyball dataset was proposed in 2016 and is limited to the domain of volleyball activity. Most algorithms achieve high accuracy in this dataset in which the best accuracy currently is 94.4%[21]. It will be worth studying whether the improvement obtained from current methods can scale up or are just the results of parameter regularization. Eventually, the dataset characterized by real-world challenging scenarios is significant for promoting research progress. Detailed annotation of various attributes such as densely actor bounding boxes or human poses may provide researchers a different perspective to solve the problem.

6 Conclusions

In this paper, we present a complete review of state-of-the-art techniques for group activity recognition. These techniques became particularly attractive in recent years because of their promising prospects in the application of video surveillance and sports video analysis. We survey several aspects of the existing attempts including handcrafted feature design and models that benefit from deep architectures. We highlight the contributions of each method and analyze their advantages. Meanwhile, we demonstrate publicly available datasets and comparisons between different methods. Future research directions are also discussed. For beginners or researchers in this field, this survey paper can be used as a helpful guide for further research.

Acknowledgements

This work was supported by National Natural Science Foundation of China (Nos. 61976010, 61802011), Beijing Postdoctoral Research Foundation (No. ZZ2019-63), Beijing excellent young talent cultivation project (No. 20170000 20124G075) and "Ri xin" Training Programme Foundation for the Talents by Beijing University of Technology.

References

[1] J. M. Chaquet, E. J. Carmona, A. Fernández-Caballero. A survey of video datasets for human action and activity recognition. *Computer Vision and Image Understanding*, vol. 117, no. 6, pp. 663–659, 2013. DOI: 10.1016/j.cviu.2013.01.013.

[2] S. Herath, M. Harandi, F. Porikli. Going deeper into action recognition: A survey. *Image and Vision Computing*, vol. 60, pp. 4–21, 2017. DOI: 10.1016/j.imavis.2017.01.010.

[3] G. C. Cheng, Y. W. Wan, A. N. Saudagar, K. Namuduri, B. P. Buckles. Advances in human action recognition: A survey, [Online], Available: https://arxiv.org/abs/1501.05964, 2015.

[4] Y. Kong, Y. Fu. Human action recognition and prediction: A survey, [Online], Available: https://arxiv.org/abs/1806.11230, 2018.

[5] C. Fauzi, S. Sulistyo. A survey of group activity recognition in smart building. In *Proceedings of International Conference on Signals and Systems*, IEEE, Bali, India, pp. 13–19, 2018. DOI: 10.1109/ICSIGSYS.2018.8372651.

[6] J. K. Aggarwal, M. S. Ryoo. Human activity analysis: A review. *ACM Computing Surveys*, vol. 43, no. 3, Article number 16, 2011. DOI: 10.1145/1922649.1922653.

[7] S. A. Vahora, N. C. Chauhan. A comprehensive study of group activity recognition methods in video. *Indian Journal of Science and Technology*, vol. 10, no. 23, 2017. DOI: 10.17485/ijst/2017/v10i23/113996.

[8] S. Blunsden, R. B. Fisher. The BEHAVE video dataset: Ground truthed video for multi-person behavior classification. *Annals of the BMVA*, vol. 2010, no. 4, pp. 1–12, 2010.

[9] W. Choi, K. Shahid, S. Savarese. 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*, IEEE, Kyoto, Japan, pp. 1282–1289, 2009. DOI: 10.1109/ICCVW.2009.5457461.

[10] M. S. Ibrahim, S. Muralidharan, Z. W. Deng, A. Vahdat, G. Mori. A hierarchical deep temporal model for group activity recognition. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Las Vegas, USA, pp. 1971–1980, 2016. DOI: 10.1109/CVPR.2016.217.

[11] V. Ramanathan, J. Huang, S. Abu-El-Haija, A. Gorban, K. Murphy, F. F. Li. Detecting events and key actors in multi-person videos. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Las Vegas, USA, pp. 3043–3053, 2016. DOI: 10.1109/CVPR.2016.332.

[12] Z. W. Cheng, L. Qin, Q. M. Huang, S. Q. Jiang, Q. Tian. Group activity recognition by gaussian processes estimation. In *Proceedings of the 20th International Conference on Pattern Recognition*, IEEE, Istanbul, Turkey, pp. 3228–3231, 2010. DOI: 10.1109/ICPR.2010.789.

[13] C. Zhang, X. K. Yang, W. Y. Lin, J. Zhu. Recognizing human group behaviors with multi-group causalities. In *Proceedings of IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*, IEEE, Macau, China, pp. 44–48, 2012. DOI: 10.1109/WI-IAT.2012.162.

[14] Y. Tang, Z. Wang, P. Li, J. Lu, M. Yang, J. Zhou. Mining semantics-preserving attention for group activity recognition. *In Proceedings of the 26th ACM International Conference on Multimedia*, Seoul, Republic of Korea, pp. 1283–1291, 2018. DOI: 10.1145/3240508.3240576.

[15] S. Khamis, V. I. Morariu, L. S. Davis. Combining per-frame and per-track cues for multi-person action recognition. In *Proceedings of the 12th European Conference on Computer Vision*, Springer, Florence, Italy, pp. 116–129, 2012. DOI: 10.1007/978-3-642-33718-5_9.

[16] M. R. Amer, P. Lei, S. Todorovic. Hirf: Hierarchical random field for collective activity recognition in videos. In *Proceedings of the 13th European Conference on Computer Vision*, Springer, Zurich, Switzerland, pp. 572–585, 2014. DOI: 10.1007/978-3-319-10599-4_37.

[17] M. R. Amer, S. Todorovic, A. Fern, S. C. Zhu. Monte Carlo tree search for scheduling activity recognition. In *Proceedings of IEEE International Conference on Computer Vision*, IEEE, Sydney, Australia, pp. 1353–1360, 2013. DOI: 10.1109/ICCV.2013.171.

[18] Z. W. Deng, A. Vahdat, H. X. Hu, G. Mori. Structure inference machines: Recurrent neural networks for analyzing relations in group activity recognition. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Las Vegas, USA, pp. 4772–4781, 2016. DOI: 10.1109/CVPR.2016.516.

[19] T. Lan, L. Sigal, G. Mori. Social roles in hierarchical models for human activity recognition. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Providence, USA, pp. 1354–1361, 2012. DOI: 10.1109/CVPR.2012.6247821.

[20] L. F. Wu, Z. Yang, J. Y. He, M. Jian, Y. W. Xu, D. Z. Xu, C. W. Chen. Ontology-based global and collective motion patterns for event classification in basketball videos. *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 7, pp. 2178–2190, 2020. DOI: 10.1109/TCSVT.2019.2912529.

[21] K. Gavrilyuk, R. Sanford, M. Javan, C. G. M. Snoek. Actor-transformers for group activity recognition, [Online], Available: https://arxiv.org/abs/2003.12737, 2020.

[22] C. Zalluhoglu, N. Ikizler-Cinbis. Collective sports: A Multi-task dataset for collective activity recognition. *Image and Vision Computing*, vol. 94, Article number 103870, 2020. DOI: 10.1016/j.imavis.2020.103870.

[23] R. Yan, L. X. Xie, J. H. Tang, X. B. Shu, Q. Tian. Social adaptive module for weakly-supervised group activity recognition, [Online], Available: https://arxiv.org/abs/2007.09470, 2020.

[24] B. B. Ni, S. C. Yan, A. Kassim. Recognizing human group activities with localized causalities. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Miami, USA, pp. 1470–1477, 2009. DOI: 10.1109/CVPR.2009.5206853.

[25] W. Choi, K. Shahid, S. Savarese. Learning context for collective activity recognition. In *Proceedings of the 24th IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Colorado, USA, pp. 3273–3280, 2011. DOI: 10.1109/CVPR.2011.5995707.

[26] W. Choi, S. Savarese. A unified framework for multi-target tracking and collective activity recognition. In *Proceedings of the 12th European Conference on Computer Vision*, Springer, Florence, Italy, pp. 215–230, 2012. DOI: 10.1007/978-3-642-33765-9_16.

[27] M. R. Amer, D. Xie, M. T. Zhao, S. Todorovic, S. C. Zhu. Cost-sensitive top-down/bottom-up inference for multiscale activity recognition. In *Proceedings of the 12th European Conference on Computer Vision*, Springer, Florence, Italy, pp. 187–200, 2012. DOI: 10.1007/978-3-642-33765-9_14.

[28] T. Lan, Y. Wang, W. L. Yang, S. N. Robinovitch. G. Mori. Discriminative latent models for recognizing contextual group activities. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 8, pp. 1549–1562, 2011. DOI: 10.1109/TPAMI.2011.228.

[29] K. N. Tran, A. Gala, I. A. Kakadiaris, S. K. Shah. Activity analysis in crowded environments using social cues for group discovery and human interaction modeling. *Pattern Recognition Letters*, vol. 44, pp. 49–57, 2014. DOI: 10.1016/j.patrec.2013.09.015.

[30] Z. W. Cheng, L. Qin, Q. M. Huang, S. C. Yan, Q. Tian. Recognizing human group action by layered model with multiple cues. *Neurocomputing*, vol. 136, pp. 124–135, 2014. DOI: 10.1016/j.neucom.2014.01.019.

[31] T. Lan, Y. Wang, G. Mori, S. N. Robinovitch. Retrieving actions in group contexts. In *Proceedings of European Conference on Computer Vision*, Springer, Heraklion, Greece, pp. 181–194, 2010. DOI: 10.1007/978-3-642-35749-7_14.

[32] T. Kaneko, M. Shimosaka, S. Odashima, R. Fukui, T. Sato. Viewpoint invariant collective activity recognition with relative action context. In *Proceedings of European Conference on Computer Vision*, Springer, Florence, Italy, pp. 253–262, 2012. DOI: 10.1007/978-3-642-33885-4_26.

[33] M. Nabi, A. Del Bue, V. Murino. Temporal poselets for collective activity detection and recognition. In *Proceedings of IEEE International Conference on Computer Vision Workshops*, IEEE, Sydney, Australia, pp. 500–507, 2013. DOI: 10.1109/ICCVW.2013.71.

[34] L. Lan, Y. Wang, W. L. Yang, G. Mori. Beyond actions: Discriminative models for contextual group activities. In *Proceedings of the 24th Annual Conference on Neural Information Processing Systems*, Vancouver, Canada, pp. 1216–122, 2010.

[35] X. B. Chang, W. S. Zheng, J. G. Zhang. Learning person–person interaction in collective activity recognition. *IEEE Transactions on Image Processing*, vol. 24, no. 6, pp. 1905–1918, 2015. DOI: 10.1109/tip.2015.2409564.

[36] H. Hajimirsadeghi, W. Yan, A. Vahdat, G. Mori. Visual recognition by counting instances: A multi-instance cardinality potential kernel. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Boston, USA, pp. 2596–2605, 2015. DOI: 10.1109/CVPR.2015.7298875.

[37] N. Vaswani, A. R. Chowdhury, R. Chellappa. Activity recognition using the dynamics of the configuration of interacting objects. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, IEEE, Madison, USA, 2003.

[38] S. M. Khan, M. Shah. Detecting group activities using rigidity of formation. In *Proceedings of the 13th annual ACM International Conference on Multimedia*, ACM, Singapore, pp. 403–406, 2005. DOI: 10.1145/1101149.1101237.

[39] Y. Zhou, B. B. Ni, S. C. Yan, T. S. Huang. Recognizing pair-activities by causality analysis. *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 1, Article number 5, 2011. DOI: 10.1145/1889681.1889686.

[40] Y. M. Zhang, W. N. Ge, M. C. Chang, X. M. Liu. Group context learning for event recognition. In *Proceedings of IEEE Workshop on the Applications of Computer Vision*, IEEE, Breckenridge, USA, pp. 249–255, 2012. DOI: 10.1109/WACV.2012.6163009.

[41] Y. F. Yin, G. Yang, M. J. Xu, H. Man. Small group human activity recognition. In *Proceedings of the 19th IEEE International Conference on Image Processing*, IEEE, Orlando, USA, pp. 2709–2712, 2012. DOI: 10.1109/ICIP.2012.6467458.

[42] J. Azorin-Lopez, M. Saval-Calvo, A. Fuster-Guillo, J. Garcia-Rodriguez, M. Cazorla, M. T. Signes-Pont. Group activity description and recognition based on trajectory analysis and neural networks. In *Proceedings of International Joint Conference on Neural Networks*, IEEE, Vancouver, Canada, pp. 1585–1592, 2016. DOI: 10.1109/IJCNN.2016.7727387.

[43] Y. J. Kim, N. G. Cho, S. W. Lee. Group activity recognition with group interaction zone. In *Proceedings of the 22nd International Conference on Pattern Recognition*, IEEE, Stockholm, Sweden, pp. 3517–3521, 2014. DOI: 10.1109/ICPR.2014.605.

[44] L. Sun, H. Z. Ai, S. H. Lao. Localizing activity groups in videos. *Computer Vision and Image Understanding*, vol. 144, pp. 144–154, 2016. DOI: 10.1016/j.cviu.2015.10.009.

[45] M. C. Chang, N. Krahnstoever, S. Lim, T. Yu. Group level activity recognition in crowded environments across multiple cameras. In *Proceedings of the 7th IEEE International Conference on Advanced Video and Signal Based Surveillance*, IEEE, Boston, USA, pp. 56–63, 2010. DOI: 10.1109/AVSS.2010.65.

[46] Z. J. Zha, H. W. Zhang, M. Wang, H. B. Luan, T. S. Chua. Detecting group activities with multi-camera context. *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 23, no. 5, pp. 856–869, 2013. DOI: 10.1109/TCSVT.2012.2226526.

[47] D. Zhang, D. Gatica-Perez, S. Bengio, I. McCowan. Modeling individual and group actions in meetings with layered HMMs. *IEEE Transactions on Multimedia*, vol. 8, no. 3, pp. 509–520, 2006. DOI: 10.1109/tmm.2006.870735.

[48] P. Dai, H. J. Di, L. G. Dong, L. M. Tao, G. Y. Xu. Group interaction analysis in dynamic context. *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 38, no. 1, pp. 275–282, 2008. DOI: 10.1109/TSMCB.2007.909939.

[49] M. R. Amer, S. Todorovic. A chains model for localizing participants of group activities in videos. In *Proceedings of International Conference on Computer Vision*, IEEE, Barcelona, Spain, pp. 786–793, 2011. DOI: 10.1109/ICCV.2011.6126317.

[50] T. Kaneko, M. Shimosaka, S. Odashima, R. Fukui, T. Sato. Consistent collective activity recognition with fully connected CRFs. In *Proceedings of the 21st International Conference on Pattern Recognition*, IEEE, Tsukuba, Japan, pp. 2792–2795, 2012.

[51] C. Y. Zhao, J. Q. Wang, H. Q. Lu. Learning discriminative context models for concurrent collective activity recognition. *Multimedia Tools and Applications*, vol. 76, no. 5, pp. 7401–7420, 2017. DOI: 10.1007/s11042-016-3393-3.

[52] T. Lan, L. Chen, Z. W. Deng, G. T. Zhou, G. Mori. Learning action primitives for multi-level video event understanding. In *Proceedings of European Conference on Computer Vision*, Springer, Zurich, Switzerland, pp. 95–110, 2014. DOI: 10.1007/978-3-319-16199-0_7.

[53] Z. Zhou, K. Li, X. J. He, M. M. Li. A generative model for recognizing mixed group activities in still images. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence*, IJCAI/AAAI Press, New York, USA, pp. 3654–3660, 2015.

[54] A. Krizhevsky, I. Sutskever, G. E. Hinton. ImageNet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems*, Lake Tahoe, USA, pp. 1106–1114, 2012.

[55] B. Zhao, J. S. Feng, X. Wu, S. C. Yan. A survey on deep learning-based fine-grained object classification and semantic segmentation. *International Journal of Automation and Computing*, vol. 14, no. 2, pp. 119–135, 2017. DOI: 10.1007/s11633-017-1053-3.

[56] V. K. Ha, J. C. Ren, X. Y. Xu, S. Zhao, G. Xie, V. Masero, A. Hussain. Deep learning based single image super-resolution: A survey. *International Journal of Automation and Computing*, vol. 16, no. 4, pp. 413–426, 2019. DOI: 10.1007/s11633-019-1183-x.

[57] K. Simonyan, A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *Proceedings of the 27th International Conference on Neural Information Processing Systems*, Montreal, Canada, pp. 568–576, 2014.

[58] M. S. Wang, B. B. Ni, X. K. Yang. Recurrent modeling of interaction context for collective activity recognition. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Honolulu, USA, pp. 7408–7416, 2017. DOI: 10.1109/CVPR.2017.783.

[59] T. M. Shu, S. Todorovic, S. C. Zhu. CERN: Confidence-energy recurrent network for group activity recognition. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Honolulu, USA, pp. 4255–4263, 2017. DOI: 10.1109/CVPR.2017.453.

[60] X. Li, M. C. Chuah. SBGAR: Semantics based group activity recognition. In *Proceedings of IEEE International Conference on Computer Vision*, IEEE, Venice, Italy, pp. 2895–2904, 2017. DOI: 10.1109/ICCV.2017.313.

[61] M. S. Qi, J. Qin, A. N. Li, Y. H. Wang, J. B. Luo, L. Van Gool. StagNet: An attentive semantic RNN for group activity recognition. In *Proceedings of the 15th European*

Conference on Computer Vision, Springer, Munich, Germany, pp. 104–120, 2018. DOI: 10.1007/978-3-030-01249-6_7.

[62] M. S. Ibrahim, G. Mori. Hierarchical relational networks for group activity recognition and retrieval. In *Proceedings of the 15th European Conference on Computer Vision*, Springer, Munich, Germany, pp. 742–758, 2018. DOI: 10.1007/978-3-030-01219-9_44.

[63] S. M. Azar, M. G. Atigh, A. Nickabadi, A. Alahi. Convolutional relational machine for group activity recognition. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Long Beach, USA, pp. 7884–7893, 2019. DOI: 10.1109/CVPR.2019.00808.

[64] J. C. Wu, L. M. Wang, L. Wang, J. Guo, G. S. Wu. Learning actor relation graphs for group activity recognition. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Long Beach, USA, pp. 9956–9966, 2019. DOI: 10.1109/CVPR.2019.01020.

[65] G. Y. Hu, B. Cui, Y. He, S. Yu. Progressive relation learning for group activity recognition, [Online], Available: https: //arxiv.org/abs/1908.02948, 2019.

[66] R. Yan, J. H. Tang, X. B. Shu, Z. C. Li, Q. Tian. Participation-contributed temporal dynamic model for group activity recognition. In *Proceedings of the 26th ACM International Conference on Multimedia*, ACM, Seoul, Republic of Korea, pp. 1292–1300, 2018. DOI: 10.1145/3240508.3240572.

[67] L. H. Lu, Y. Lu, R. Z. Yu, H. J. Di, L. Zhang, S. Z. Wang. GAIM: Graph attention interaction model for collective activity recognition. *IEEE Transactions on Multimedia*, vol. 22, no. 2, pp. 524–539, 2020. DOI: 10.1109/TMM.2019.2930344.

[68] T. Bagautdinov, A. Alahi, F. Fleuret, P. Fua, S. Savarese. Social scene understanding: End-to-end multi-person action localization and collective activity recognition. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Honolulu, USA, pp. 4325–4334, 2017. DOI: 10.1109/CVPR.2017.365.

[69] P. Z. Zhang, Y. Y. Tang, J. F. Hu, W. S. Zheng. Fast collective activity recognition under weak supervision. *IEEE Transactions on Image Processing*, vol. 29, pp. 29–43, 2019. DOI: 10.1109/TIP.2019.2918725.

[70] S. Hochreiter, J. Schmidhuber. Long short-term memory. *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997. DOI: 10.1007/978-3-642-24797-2_4.

[71] A. Graves, N. Jaitly, A. R. Mohamed. Hybrid speech recognition with deep bidirectional LSTM. In *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding*, IEEE, Olomouc, Czech Republic, pp. 273–278, 2013. DOI: 10.1109/ASRU.2013.6707742.

[72] K. Xu, J. L. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. S. Zemel, Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the 32nd International Conference on Machine Learning*, Lille, France, pp. 2048–2057, 2015.

[73] P. S. Kim, D. G. Lee, S. W. Lee. Discriminative context learning with gated recurrent unit for group activity recognition. *Pattern Recognition*, vol. 76, pp. 149–161, 2018. DOI: 10.1016/j.patcog.2017.10.037.

[74] H. Gammulle, S. Denman, S. Sridharan, C. Fookes. Multi-level sequence GAN for group activity recognition. In *Proceedings of the 14th Asian Conference on Computer Vision*, Springer, Perth, Australia, pp. 331–346, 2018. DOI: 10.1007/978-3-030-20887-5_21.

[75] L. F. Wu, J. Y. He, M. Jian, S. Y. Liu, Y. W. Xu. Global motion pattern based event recognition in multi-person videos. In *Proceedings of the 2nd CCF Chinese Conference on Computer Vision*, Springer, Tianjin, China, pp. 667–676, 2017. DOI: 10.1007/978-981-10-7305-2_56.

[76] X. B. Shu, L. Y. Zhang, Y. L. Sun, J. H. Tang. Host-parasite: Graph LSTM-in-LSTM for group activity recognition. *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–12, 2020. DOI: 10.1109/TNNLS.2020.2978942.

[77] Z. W. Deng, M. Y. Zhai, L. Chen, Y. H. Liu, S. Muralidharan, M. J. Roshtkhari, G. Mori. Deep structured models for group activity recognition. In *Proceedings of British Machine Vision Conference*, Swansea, UK, pp. 179.1–179.12, 2015. DOI: 10.5244/C.29.179.

[78] Y. Y. Tang, P. Z. Zhang, J. F. Hu, W. S. Zheng. Latent embeddings for collective activity recognition. In *Proceedings of the 14th IEEE International Conference on Advanced Video and Signal Based Surveillance*, IEEE, Lecce, Italy, 2017. DOI: 10.1109/AVSS.2017.8078522.

[79] D. Z. Xu, H. Fu, L. F. Wu, M. Jian, D. Wang, X. Liu. Group activity recognition by using effective multiple modality relation representation with temporal-spatial attention. *IEEE Access*, vol. 8, pp. 65689–65698, 2020. DOI: 10.1109/ACCESS.2020.2979742.

[80] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, L. Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, Long Beach, USA, pp. 6000–6010, 2017.

[81] M. Ehsanpour, A. Abedin, F. Saleh, J. Shi, I. Reid, H. Rezatofighi. Joint learning of social groups, individuals action and sub-group activities in videos, [Online], Available: https://arxiv.org/abs/2007.02632, 2020.

[82] S. A. Vahora, N. C. Chauhan. Deep neural network model for group activity recognition using contextual relationship. *Engineering Science and Technology, an International Journal*, vol. 22, no. 1, pp. 47–54, 2019. DOI: 10.1016/j.jestch.2018.08.010.

[83] J. C. Liu, C. X. Wang, Y. T. Gong, H. Xue. Deep fully connected model for collective activity recognition. *IEEE Access*, vol. 7, pp. 104308–104314, 2019. DOI: 10.1109/ACCESS.2019.2929684.

[84] S. E. Wei, V. Ramakrishna, T. Kanade, Y. Sheikh. Convolutional pose machines. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Las Vegas, USA, pp. 4724–4732, 2016. DOI: 10.1109/CVPR.2016.511.

[85] T. N. Kipf, M. Welling. Semi-supervised classification with graph convolutional networks, [Online], Available: https://arxiv.org/abs/1609.02907, 2017.

[86] J. Y. Gao, T. Z. Zhang, C. S. Xu. Graph convolutional tracking. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Long Beach,USA,pp. 4644–4654,2019.DOI:10.1109/CVPR.2019.00478.

[87] S. J. Yan, Y. J. Xiong, D. H. Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, New Orleans, USA, pp. 7444–7452, 2018.

[88] X. L. Wang, A. Gupta. Videos as space-time region graphs. In *Proceedings of the 15th European Conference on Computer Vision*, Springer, Munich, Germany, pp. 413–431, 2018. DOI: 10.1007/978-3-030-01228-1_25.

[89] J. H. Tang, X. B. Shu, R. Yan, L. Y. Zhang. Coherence constrained graph LSTM for group activity recognition.

IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019. DOI: 10.1109/TPAMI.2019.2928540.

[90] Y. S. Tang, J. W. Lu, Z. Wang, M. Yang, J. Zhou. Learning semantics-preserving attention and contextual interaction for group activity recognition. *IEEE Transactions on Image Processing*, vol. 28, no. 10, pp. 4997–5012, 2019. DOI: 10.1109/TIP.2019.2914577.

[91] L. H. Lu, H. J. Di, Y. Lu, L. Zhang, S. Z. Wang. A two-level attention-based interaction model for multi-person activity recognition. *Neurocomputing*, vol. 322, pp. 195–205, 2018. DOI: 10.1016/j.neucom.2018.09.060.

[92] L. H. Lu, H. J. Di, Y. Lu, L. Zhang, S. Z. Wang. Spatio-temporal attention mechanisms based model for collective activity recognition. *Signal Processing*: *Image Communication*, vol. 74, pp. 162–174, 2019. DOI: 10.1016/j.image.2019.02.012.

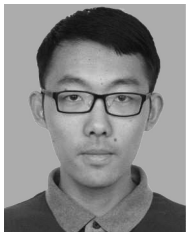
[93] N. F. Zhuang, T. Yusufu, J. Ye, K. A. Hua. Group activity recognition with differential recurrent convolutional neural networks. In *Proceedings of the 12th IEEE International Conference on Automatic Face & Gesture Recognition*, IEEE, Washington, USA, pp. 526–531, 2017. DOI: 10.1109/FG.2017.70.

Li-Fang Wu received the B. Eng. degree in radio technology and M.Eng. degree in metal material and heat treatment from Beijing University of Technology (BJUT), China in 1991 and 1994, respectively, and the Ph. D. degree in pattern recognition and intelligent system from BJUT in 2003. She is a professor with Faculty of Information Technology, Beijing University of Technology, China. She has published over 100 referred technical papers in international journals and conferences of image/video processing and pattern recognition. She is a senior member of the China Computer Federation.

Her research interests include image/video analysis and understanding, social media computing, intelligent 3D printing and face presentation attack detection.

E-mail: lfwu@bjut.edu.cn

ORCID iD: 0000-0002-7209-0215



Qi Wang received the B. Sc. degree in electronic information engineering from Beijing University of Technology, China in 2018. He is currently a master student in information and communication engineering at College of Information and Communication Engineering, Beijing University of Technology, China.

His research interests include group activity recognition, computer vision and image processing.

E-mail: qiwang@emails.bjut.edu.cn

Meng Jian received the B. Sc. degree in electronic information science and technology and Ph. D. degree in pattern recognition and information system from Xidian University, China in 2010 and 2015, respectively. She is currently an associate professor with Faculty of Information Technology, Beijing University of Technology, China. She is also a Research Scholar with School of Computing, National University of Singapore, Singapore, from November 2018 to November 2019. She has been awarded Beijing Excellent Young Talent in 2017 and "Ri xin" Talents of Beijing University of Technology in 2018.

Her research interests include pattern recognition, image understanding and social media computing.

E-mail: jianmeng648@163.com(Corresponding author)

ORCID iD: 0000-0001-5659-5128

Yu Qiao received Ph. D. degree in information system from University of Electro-Communications, Japan in 2006. He is currently a professor and director of Institute of Advanced Computing and Digital Engineering, with the Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences. He has been a Japan Science Promotion Society Fellow and a project assistant professor with The University of Tokyo, Japan from 2007 to 2010. He has authored over 180 articles in journals and conferences, including PAMI, IJCV, TIP, ICCV, CVPR, ECCV and AAAI, with h-index 52. He was a recipient of the Lu Jiaxi Young Researcher Award from the Chinese Academy of Sciences in 2012, and the first class award on technological invention from Guangdong provincial government in 2019. He was the winner of video classification task in the ActivityNet Large Scale Activity Recognition Challenge 2016 and the first Runner-up of scene recognition task in the ImageNet Large Scale Visual Recognition Challenge 2015.

His research interests include computer vision, deep learning and intelligent robots.

E-mail: yu.qiao@siat.ac.cn

Bo-Xuan Zhao received the B. Sc. degree in electronic information engineering from Beijing University of Technology, China in 2019, and is currently a master student in electronic and communications engineering at Beijing University of Technology, China.

His research interests include target tracking and motion trajectory description.

E-mail: zhaoboxuan@emails.bjut.edu.cn