RIT-18: A Novel Dataset for Compositional Group Activity Understanding

Junwen Chen Haiting Hao Hanbin Hong Yu Kong
Golisano College of Computing and Information Sciences
Rochester Institute of Technology
Rochester, NY 14623, USA
{jc1088, hh7702, hh9665, Yu.Kong}@rit.edu

Abstract

Group activity understanding is a challenging task as multiple people are involved, and their relations may vary over time. Currently, the literature of group activity is limited to group activity recognition, because videos are trimmed in very short duration and focus on a single activity. This slows down the progress in the group activity domain. In this paper, we propose a new large-scale untrimmed compositional group activity dataset RIT-18 based on the volleyball games captured from YouTube. Each clip in our dataset depicts an entire rally which spans the duration from serve to a point being scored. Comprehensive annotations including group activity labels, temporal boundaries of activities, key persons, and winning teams are provided. We describe group activity recognition, future activity anticipation, and rally-level winner prediction challenges, and evaluate several baseline methods over these challenges. We report their performance on our dataset and demonstrate further efforts need to be made. The dataset is available at `https://pht180.rit.edu/actionlab/rit-18`.

1. Introduction

Group activity is gaining interest owing to its broad applications in surveillance [5], crime prevention [22] and autonomous driving [11]. The goal is to understand the activity performed by multiple people that belong to a group. It is of great significance, as in most of the outdoor surveillance scenes, there is more than one person. Action understanding for single people or two people has been well explored and diversified in various extended tasks, including future anticipation [10, 24], human-object-interaction [23, 17] and temporal activity localization [26]. But for group activity, it is less explored. Currently, Volleyball Dataset [14] and Collective Activity Dataset [5] are two popularly used datasets for group activity recognition. Based on these, many group activity recognition methods [14, 18, 27, 21] have been developed and especially with the prosperity of relation learning [13, 25], and the recognition performance has been boosted. However, these datasets are either too short in duration or too small in size, i.e. only 41 frames trimmed for each activity clip in Volleyball Dataset. The good performance on these datasets is limited to recognition under constrained and can be hardly generalized to the real challenging scenes. More importantly, because of the short-duration and trimmed videos, the existing dataset cannot benchmark more challenging action-related tasks, e.g. activity anticipation [10]. Activity anticipation requires to predict activity labels given only the beginning part of the video. It mimics the real urgent scenario where the intelligent system needs to predict activity before it ends, in order to avoid high-risk outcomes. This also applies to group activity in surveillance scenes. But the existing trimmed videos do not contain the progression of activities and hence cannot evaluate the future anticipation. Besides, for activity understanding, it is not enough to simply infer the activity class. It is of great significance if the goal and intention behind the activity can be predicted. To this end, for group activity in a volleyball game, predicting

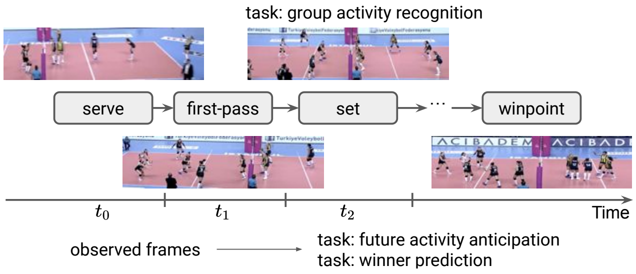


Figure 1. In our RIT-18 dataset, each clip contains an entire rally that starts from a serve and goes through many hits until a point being scored. The activity class, temporal boundary and winner team of the rally are crowd-sourced, which facilitate the research for multiple tasks including activity recognition, future anticipation and winner prediction.

“which team will score the rally” should be a valuable task.

In this paper, we propose a novel group activity dataset, named as RIT-18, which extends the existing Volleyball Dataset [14]. The dataset is collected from the 51 volleyball game videos on YouTube. We extract a rally which starts from a service and goes through many hits until a point being scored, as a clip. For each clip, the activity classes and temporal boundaries of group activities are annotated. Thus, it can be considered as a compositional activity describing the entire rally. In total, we have 18 group activity classes, which is more diverse than Volleyball Dataset [14]. The temporal boundaries of group activities are provided, which enables the evaluation on various activity anticipation tasks, e.g. “What will happen in the next second?” and “What’s the next activity following the current one?”. Besides, the winner team of a rally is also annotated, which enables a longer-term prediction of winner. “Predicting the winner” is challenging due to many uncertain factors in the progress of a sports game. It requires a prediction model that learns to understand the group intentions behind their motions. Thus, it cannot be directly considered as a binary classification task. Lastly, recent work [27, 9, 19, 2, 12, 4, 16] on group activity recognition assumes only part of people’s actions influence the group activity, while others are irrelevant. Inspired by this, we labeled the positions and actions of key persons of each activity by who is touching the ball, which provides a quantitative evaluation of individual contribution in a group activity.

To sum up, the novel RIT-18 dataset offers comprehensive evaluations of group activity understanding via several challenges including group activity recognition, group activity anticipation and winner prediction.

2. Related Dataset

We compare the collected RIT-18 dataset with several popular group activity and interaction datasets in Table. 1. The human-object interaction is excluded because it is a different research question.

Some existing datasets [15] focus on the action performed by two people, also called *interactions*. Previous work proposes solutions for both the interaction recognition [15] and interaction anticipation [28]. For activity performed by a group of people, a few datasets [5, 8, 14] have been proposed. Group activity datasets have been built in both surveillance [5] and sports scenes [14]. In recent year, a large amount of methods [1, 21, 14, 3] have achieved good performance in group activity recognition tasks, mainly owing to a good relational modeling [25, 13] and suppressing the irrelevant people [18, 27]. However, the datasets on which they conduct experiments are either of small size, i.e. only 44 videos in collective activity dataset, or of short duration, i.e. only 41 frames per clip trimmed for each activity in volleyball dataset. Thus, to better analyze group activity, a dataset with untrimmed group activity is desired, so that more challenging tasks can be investigated.

Our RIT-18 dataset follows Volleyball dataset [14], but extends it to the settings with large-scale and long duration. In Volleyball dataset [14], each clip contains a short duration of a single volleyball activity such as spiking. In RIT-18 dataset, each clip describes a rally that starts from serving, goes through many hits and ends with the ball landing in or landing out, which is a compositional activity. Using this dataset, both the group activity recognition and anticipation can be evaluated.

Winner Prediction Winner prediction aims at predicting which team will score only given the beginning frames of the compositional activity. Winner prediction is driven by a general intention of a sports game. Tt can be considered as goal-based task, different from the label-based classification task such as activity recognition. Another challenge of winner prediction is a longer-term prediction than activity anticipation and the temporal lengths of a clip are various. Many uncertain factors might occur in sports games. Because of these challenges, only very few work has been proposed for winner prediction. Previously, Decroos et al. [6, 7] presents a model to predict short-term scoring and conceding probabilities at any moment in a soccer game. But the prediction is based on the summarized language description of the game. To our best knowledge, there is no existing work that predicts the winner from the visual signals, even if visual signals of sports game are ubiquitous online. In our dataset, we annotate the winner team of each rally and offer a benchmark for long-term winner prediction.

3. The RIT-l8 Activity Dataset

3.1. Data Collection

We collected video clips from 51 volleyball games on YouTube. To increase diversity, we select volleyball games at different professional levels, including Olympics, World Championship, World League, and NCAA . The games are almost half-and-half in gender. Around 88% of the games in the dataset were held after 2016. Figure. 2 shows the detailed summary of the dataset distribution.

For each game, we divided it into many clips, each of which contains a rally from serving to a point being scored. We filtered out very short clips, such as “ace”. The length of clip varies from 4.04 seconds to 50.50 seconds, depending on the actual time duration of the rally. On average, the length of each clip is 9.09 seconds. In total, we collect 1530 clips.

In contrast to Volleyball Dataset [14], video clips in our dataset are longer and diverse as it consists of the compositional activities of entire rally, while [14] only ranges over a trimmed activity.

Table 1. A list of the existing group activity recognition datasets. Comp. and Anno. mean compositional activity and annotations, respectively. Bala. means if the instance numbers of activity labels are balanced. The annotations of each dataset are described in the Anno. column, where T for tracklets, W for the winner, A for activity, and B for the temporal boundary of an activity. Our RIT-18 dataset provides comprehensive annotations for evaluating group activity understanding methods.

Dataset	Year	Comp.	Anno.	Bala.	#Clips	#Activities	#People	Env.	#Frames/clip
Collective Activity [5]	2009	No	TA	Yes	44	7	multiple	Surveillance	570
UT-interaction [20]	2010	No	A	Yes	60	6	2	Surveillance	≈ 100
BIT [15]	2012	No	A	Yes	400	8	2	Surveillance	≈ 100
Volleyball [14]	2016	No	TA	Yes	4830	8	multiple	Sports	41
RIT-18	2020	Yes	TWAB	No	1530	18	multiple	Sports	298

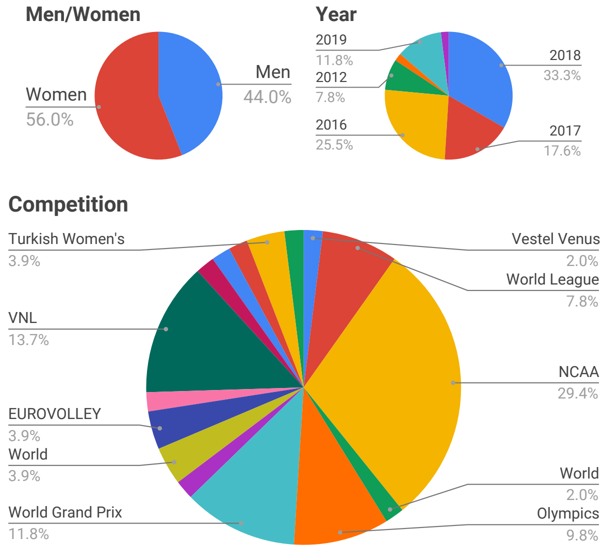


Figure 2. The statistics of RIT-18 dataset: Genders, Years, and Competitions.

3.2. Annotations

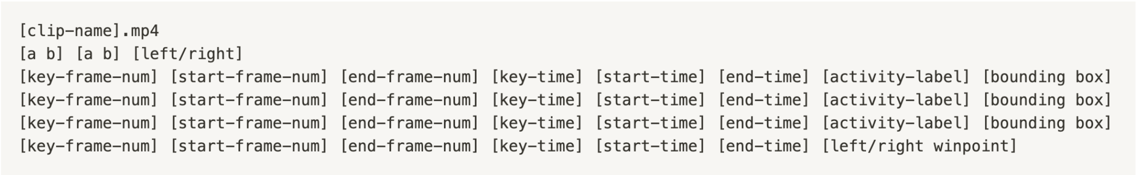
Each video clip in our dataset contains an entire rally starting from serving to a point being scored. It contains multiple group activities, including serve, firstpass, set, spike, etc. We have 18 group activity classes, which are more than [14] where only 8 classes are annotated. We calculate the instances number of activities in Table. 2. Table. 2 shows that the numbers of different activity labels are diversified. For example, the *shot* activity samples are far less than *l-firstpass*. This is because some activities happen when the players are in a very passive situation, leading to the solutions of rare group activity recognition.

For each clip, we annotate temporal boundaries of each group activity, which can be used in temporal group activity localization and group activity anticipation tasks. Currently, there is no benchmark for addressing these tasks that can be applied to video retrieval. We also label the bounding box of the player who is touching the ball in the middle frame of each activity. Based on the annotation, we can explicitly investigate the contribution of individuals to a group. Note that key players are not labeled for "winpoint" activity, because it is at the moment of ball landing "in" or landing "out", in which nobody is touching the ball. Moreover, for each clip, we also provide annotations of the winner in the rally, in order to encourage a solution for long-term goal-based prediction of group activity.

In total, we annotate 12035 frames from 1530 clips (51 games × 30 rallies per game) with 18 group activity labels. The overall labeling structure of our compositional group activity is shown in Figure 3. The accumulated scores of 51 games are also recorded in the dataset, which points to a longer-term game-level winner prediction.

Table 2. Instance numbers of the 18 group activity classes in RIT-18 dataset.

Activity Label	# of Instances
l-serve	774
l-firstpass	1395
l-set	1270
l-spike	1012
l-volley	136
l-drop	243
l-shot	23
l-block	393
l-winpoint	764
r-serve	756
r-firstpass	1395
r-set	1266
r-spike	1014
r-volley	143
r-drop	235
r-shot	23
r-block	427
r-winpoint	766
Total	12035



```
[clip-name].mp4
[a b] [a b] [left/right]
[key-frame-num] [start-frame-num] [end-frame-num] [key-time] [start-time] [end-time] [activity-label] [bounding box]
[key-frame-num] [start-frame-num] [end-frame-num] [key-time] [start-time] [end-time] [activity-label] [bounding box]
[key-frame-num] [start-frame-num] [end-frame-num] [key-time] [start-time] [end-time] [activity-label] [bounding box]
[key-frame-num] [start-frame-num] [end-frame-num] [key-time] [start-time] [end-time] [left/right winpoint]
```

Figure 3. The annotation format of a rally/clip in RIT-18 dataset. For each clip, we record the current score and the winner (left / right is based on the position at the beginning of the match). The temporal boundary of each activity is labeled by frame-id and seconds. The bounding box of the key person who is touching the ball is annotated on the middle frame of the activity.

Table 3. Group activity recognition results on RIT-18 dataset. We show the accuracy (%) for each class and mean accuracy (%). Existing state-of-the-art group activity recognition methods (SSU, ARG, HDTM) are the comparison baselines.

HDTM [14]	l-serve	l-spike	l-set	l-pass	l-block	l-win	l-drop	l-volley	l-shot
	28.95	75.92	58.33	23.73	0.0	44.30	0.0	0.0	0.0
mean	r-serve	r-spike	r-set	r-pass	r-block	r-win	r-drop	r-volley	r-shot
40.41	25.69	65.97	67.65	43.99	0.0	16.89	0.0	0.0	0.0
SSU [3]	l-serve	l-spike	l-set	l-pass	l-block	l-win	l-drop	l-volley	l-shot
	92.98	91.62	88.41	58.54	3.23	42.41	20.69	22.22	0.0
mean	r-serve	r-spike	r-set	r-pass	r-block	r-win	r-drop	r-volley	r-shot
67.51	95.41	95.81	86.40	78.69	0.0	51.35	1.45	5.56	0.0
ARG [25]	l-serve	l-spike	l-set	l-pass	l-block	l-win	l-drop	l-volley	l-shot
	95.61	87.43	88.77	60.76	8.06	60.76	10.34	36.11	0.0
mean	r-serve	r-spike	r-set	r-pass	r-block	r-win	r-drop	r-volley	r-shot
68.09	98.17	91.62	84.19	76.98	10.53	38.51	10.14	8.33	0.0

4. Baseline Results

Our dataset offers a variety of challenges including group activity recognition, future anticipation and long-term winner prediction, etc. In this paper, we provide baseline results for the three challenges. We evaluate several existing methods of group activity recognition on our benchmark, to show the difficulty and application value of our dataset. For all of the experiments, the dataset is divided into a training split with 37 videos and a test split with 14 videos.

4.1. Group Activity Recognition Challenge

Evaluation metric: Group activity recognition challenge requires to recognize the group activity given a trimmed video clip. We report the top-1 accuracy of recognition of each class.

Baselines: The state-of-the-art group activity recognition methods ARG [25] and SSU [3] are trained and tested on RIT-18 dataset. SSU takes as input the video frames and achieves both the detection of people and the group activity recognition in the same framework. ARG follows SSU but proposes to build a learnable graph to capture the relations between people.

Results: The testing results are shown in Table. 3. ARG achieved 92.14% accuracy in Volleyball Dataset [14] but only achieves 68.09% in RIT-18 dataset. This is because our group activity labels are more diverse than [14] (18 vs 8) and our volleyball games come from different professional levels ranging from Olympic games to NCAA. Table. 3 also shows that activities such as serve, spike, set are easy to be recognized, but the rare activity classes i.e. drop, volley and shot are difficult to be recognized. The rare classes mainly appear in some faulty activities or when a team is in a very negative situation. We leave the rare group activity recognition as a future direction.

4.2. Group Activity Anticipation Challenge

Evaluation metric: Group activity anticipation requires to anticipate the activity class given a partial video with an incomplete rally, which is defined as a partial observation. The observation ratio of a partial observation on an ongoing activity is defined as the percentage of the observation length over the total length of the corresponding full activity execution. In our task, the observation ratio of the activity varies and is unknown, since in some real applications, we are not able to know how much the activity has been completed. In this paper, the challenge is defined to predict the activity classes in the future 0.5 second, future 1 second and the classes of next activity following the ongoing one. We report the mean top-1 accuracy of future anticipation, given the partial observation that at different observation ratios including 10%, 40% and 70%.

Baselines: We extend the state-of-the-art group activity recognition methods ARG and SSU to achieve anticipation of 0.5s, 1s and next activity, given the partial observation.

Table 4. Group activity anticipation results on RIT-18 dataset. We show the mean accuracy (%) of prediction for future 0.5s, 1s and the next activity after the ongoing activity, given the partially observed frames of the ongoing activity. The percentage of the partial observation length divided by the total length of the corresponding activity is defined as observation ratio. In this experiment, we report the anticipation results given the ongoing activity at observation ratio 10%, 40% and 70% respectively.

Method	10%			40%			70%		
	0.5s	1s	next	0.5s	1s	next	0.5s	1s	next
ARG [25]	37.43	36.93	38.52	44.88	47.43	49.90	49.10	51.69	55.92
SSU [3]	38.06	36.64	36.43	42.53	40.52	42.62	42.62	43.72	45.21

The three anticipation scopes share the backbone pretrained on the recognition model in Sec.4.1 and the anticipation branches are independently trained.

Results: The testing results are shown in Table. 4. The accuracy of baseline methods significantly drops when observing very few frames e.g. 10%. A future direction of this challenge is how to enhance the discriminative features when given limited beginning frames.

4.3. Rally-level Winner Prediction Challenge

Winner prediction challenge is designed to benchmark the long-term goal-based prediction performance. A good winner prediction is supposed to implicitly model the general intention of players.

Evaluation metric: Given a partial observation of a clip (rally), the model needs to predict the winner team of the clip, which is a binary classification of either left or right. The observation ratio of the clip is various and unknown. We report the winner prediction results by observing the beginning 10%, 40%, 70% and 100% of each rally, defined as observation ratio below. Note that 100% is considered as winner recognition where all of the frames in the rally including winpoint are observed.

Baselines: To our best knowledge, there is no existing method developed for visual-based winner prediction. Existing group activity recognition method ARG [25] cannot be directly used for winner prediction, since it does not model the temporal context. Below we design three baselines.

B1 LSTM: We first extract frame-level features by VGG19 and apply LSTM on the sequence of frame features. The output of LSTM is used for binary classification.

B2. Extended ARG with temporal order: We extend ARG [25] by sorting the six frames from the partial observation, selected by the sparse temporal sampling strategy.

B3. Extended ARG with monotone importance: We extend ARG [25] by sorting the six frames from the partial observation and increasing the importance of the latter frames gradually, since they are closer to winpoint and should be more responsible for the prediction.

Results: According to Table. 5, *B1 LSTM* is not able to predict the winner even if given the full rally, since it only extracts features of the entire frames but does not model the relations between people. *Extended ARG B2&B3* are able to predict the winner to some extent. As approaching the end of the clip, the prediction accuracy is increasing. Monotone importance (B3) increases the accuracy of winner prediction, by comparison with sorted ARG (B2). The prediction at 10% of the rally is close to random guess, because it is a very long-duration setting and some rally takes more than 30 seconds with athletes quickly moving and changing poses.

Table 5. Winner prediction accuracy (%) on RIT-18 dataset. Winner prediction is experimented given the partially observed rally at three different observation ratios, 10%, 40%, 70% and 100% correspondingly. The observation ratio is defined by the observation length over the length of an entire rally. Three baselines are designed for evaluation.

Method	10%	40%	70%	100%
B1	51.25%	51.25%	51.25%	51.25%
B2	54.98%	58.41%	62.54%	64.94%
B3	51.89%	53.61%	67.69%	75.61%

Winner prediction is different from activity recognition, because it is a goal-based prediction while recognition is label-based classification. Two adversarial teams attempt to achieve their goal and the intention of the individual motions and group formation should be learned. An implicit quality estimation of individual physical skill and interperson coordination may benefit winner prediction, which is an open question remaining for future work.

5. Conclusion and Future work

In this paper, we propose a novel and challenging dataset to investigate group activity understanding. We collect 1530 rallies from 51 volleyball games and annotate them with group activity labels, temporal boundaries, key persons and winner teams. Baseline results on group activity recognition, anticipation and winner prediction suggest that RIT-18 dataset is challenging and current methods that overwhelmed in the existing datasets still remain limited in fully understanding the group activity. Future work is pointed as follows.

Defined challenges: The three challenges defined above can formulate a thorough understanding of group activity. Particularly, the winner prediction task is designed to evaluate a high-level understanding of the goal of players. The baseline results on the three challenges are still far from solving them with high precision. The future directions of the three challenges lie in long-term temporal modeling of multiple people, a way to recognize the rare activity class and a spatio-temporal learning of inter-person interactions.

Other possible challenges on RIT-18 dataset: In addition to the defined three tasks, our dataset has the potential to evaluate other less explored tasks.

- **Temporal group activity localization**: As temporal boundaries are provided in our untrimmed compositional activity dataset, temporal localization of group activities can be achieved. The temporal localization needs a good temporal context learning of multiple people. The diverse clip lengths raise the challenge of discovering the inter-person interactions.

- **Individual Contribution**: We provide the bounding box annotation of the player who is touching the ball in each activity, as the key person. It is of great significance to infer the contribution of individuals in forming this group activity, with applications in crime prevention and athletes valuing. The participation of individuals has been investigated in existing group activity recognition work [27, 18] during training stage, to suppress the irrelevant players. But it was not quantitatively evaluated before because of no annotations. Our key person annotations can be used to evaluate individual contributions to a group activity.

- **Game-level winner prediction**: According to Figure 3, we can achieve a longer duration prediction based on RIT-18. Since the game winner, set winner and accumulated points are all annotated, we can predict the set winner or the game winner, given some randomly selected clips of the set or the game. It is essentially a visual-based skill determination task to infer which team is better organized and more skilled.

Acknowledgement

We thank Nvidia for the GPU donation. This research is supported in part by ONR Award N00014-18-1-2875.

References

[1] Mohamed Rabie Amer, Peng Lei, and Sinisa Todorovic. Hirf: Hierarchical random field for collective activity recognition in videos. In *ECCV*, pages 572–585, 2014.

[2] Sina Mokhtarzadeh Azar, Mina Ghadimi Atigh, Ahmad Nickabadi, and Alexandre Alahi. Convolutional relational machine for group activity recognition. In *CVPR*, pages 7892–7901, 2019.

[3] Timur Bagautdinov, Alexandre Alahi, François Fleuret, Pascal Fua, and Silvio Savarese. Social scene understanding: End-to-end multi-person action localization and collective activity recognition. In *CVPR*, pages 4315–4324, 2017.

[4] Sovan Biswas and Juergen Gall. Structural recurrent neural network (srnn) for group activity analysis. In *WACV*, pages 1625–1632, 2018.

[5] Wongun Choi, Khuram Shahid, and Silvio Savarese. What are they doing?: Collective activity classification using spatio-temporal relationship among people. In *ICCV Workshops*, pages 1282–1289, 2009.

[6] Tom Decroos, Lotte Bransen, Jan Van Haaren, and Jesse Davis. Actions speak louder than goals: Valuing player actions in soccer. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1851–1861, 2019.

[7] Tom Decroos, Vladimir Dzyuba, Jan Van Haaren, and Jesse Davis. Predicting soccer highlights from spatio-temporal match event streams. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[8] Zhiwei Deng, Mengyao Zhai, Lei Chen, Yuhao Liu, Srikanth Muralidharan, Mehrsan Javan Roshtkhari, and Greg Mori. Deep structured models for group activity recognition. *arXiv preprint arXiv:1506.04191*, 2015.

[9] Harshala Gammulle, Simon Denman, Sridha Sridharan, and Clinton Fookes. Multi-level sequence gan for group activity recognition. In *Asian Conference on Computer Vision*, pages 331–346, 2018.

[10] Harshala Gammulle, Simon Denman, Sridha Sridharan, and Clinton Fookes. Predicting the future: A jointly learnt model for action anticipation. In *ICCV*, pages 5562–5571, 2019.

[11] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013.

[12] Guyue Hu, Bo Cui, Yuan He, and Shan Yu. Progressive relation learning for group activity recognition. *arXiv preprint arXiv:1908.02948*, 2019.

[13] Mostafa S Ibrahim and Greg Mori. Hierarchical relational networks for group activity recognition and retrieval. In *ECCV*, pages 721–736, 2018.

[14] Mostafa S Ibrahim, Srikanth Muralidharan, Zhiwei Deng, Arash Vahdat, and Greg Mori. A hierarchical deep temporal model for group activity recognition. In *CVPR*, pages 1971–1980, 2016.

[15] Yu Kong, Yunde Jia, and Yun Fu. Learning human interaction by interactive phrases. In *European conference on computer vision*, pages 300–313. Springer, 2012.

[16] Xin Li and Mooi Choo Chuah. Sbgar: Semantics based group activity recognition. In *ICCV*, pages 2876–2885, 2017.

[17] Joanna Materzynska, Tete Xiao, Roei Herzig, Huijuan Xu, Xiaolong Wang, and Trevor Darrell. Something-else: Compositional action recognition with spatial-temporal interaction networks. *CVPR*, 2020.

[18] Mengshi Qi, Jie Qin, Annan Li, Yunhong Wang, Jiebo Luo, and Luc Van Gool. stagnet: An attentive semantic rnn for group activity recognition. In *ECCV*, pages 101–117, 2018.

[19] Vignesh Ramanathan, Jonathan Huang, Sami Abu-El-Haija, Alexander Gorban, Kevin Murphy, and Li Fei-Fei. Detecting events and key actors in multi-person videos. In *CVPR*, pages 3043–3053, 2016.

[20] Michael S Ryoo, Chia-Chih Chen, JK Aggarwal, and Amit Roy-Chowdhury. An overview of contest on semantic description of human activities (sdha) 2010. In *International Conference on Pattern Recognition*, pages 270–285. Springer, 2010.

[21] Tianmin Shu, Sinisa Todorovic, and Song-Chun Zhu. Cern: confidence-energy recurrent network for group activity recognition. In *CVPR*, pages 5523–5531, 2017.

[22] Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6479–6488, 2018.

[23] Xiaolong Wang and Abhinav Gupta. Videos as space-time region graphs. In *ECCV*, pages 399–417, 2018.

[24] Xionghui Wang, Jian-Fang Hu, Jian-Huang Lai, Jianguo Zhang, and Wei-Shi Zheng. Progressive teacher-student learning for early action prediction. In *CVPR*, pages 3556–3565, 2019.

[25] Jianchao Wu, Limin Wang, Li Wang, Jie Guo, and Gangshan Wu. Learning actor relation graphs for group activity recognition. In *CVPR*, pages 9964–9974, 2019.

[26] Huijuan Xu, Abir Das, and Kate Saenko. R-c3d: Region convolutional 3d network for temporal activity detection. In *Proceedings of the IEEE international conference on computer vision*, pages 5783–5792, 2017.

[27] Rui Yan, Jinhui Tang, Xiangbo Shu, Zechao Li, and Qi Tian. Participation-contributed temporal dynamic model for group activity recognition. In *Multimedia*, pages 1292–1300, 2018.

[28] Yichao Yan, Bingbing Ni, and Xiaokang Yang. Predicting human interaction via relative attention model. *IJCAI*, pages 3245–3251, 2017.