17th International Conference on Machine Vision Applications (MVA)
Fully Online, July 25–27, 2021.

Group Activity Recognition Using Joint Learning of Individual Action Recognition and People Grouping

Chihiro Nakatani Kohei Sendo Norimichi Ukita
Toyota Technological Institute

Abstract

This paper proposes joint learning of individual action recognition and people grouping for improving group activity recognition. By sharing the information between two similar tasks (i.e., individual action recognition and people grouping) through joint learning, errors of these two tasks are mutually corrected. This joint learning also improves the accuracy of group activity recognition. Our proposed method is designed to consist of any individual action recognition methods as a component. The effectiveness is validated with various IAR methods. By employing existing group activity recognition methods for ensembling with the proposed method, we achieved the best performance compared to the similar SOTA group activity recognition methods.

1 Introduction

Human action recognition is one of the most important topics in computer vision. While Individual Action Recognition (IAR) has been studied actively, Group Activity Recognition (GAR) is still under development.

For GAR, graphical models that represent dependencies between random variables are often used [1, 2, 3, 4]. These models first recognize the actions of individual people and then recognize their group activity based on the individual actions and the location of those people. In such stepwise approaches shown in Fig. 1 (a), the erroneous results of IAR may adversely affect the subsequent people grouping and GAR. To reduce such errors, our proposed method, as shown in Fig. 1 (b) improves IAR and people grouping through joint learning, and finally improves GAR.

Our work provides the following three contributions:

- We empirically validate that joint learning of IAR and people grouping is effective for GAR.
- Since our joint learning can be integrated with any IAR methods, its effectiveness is validated with various IAR methods.
- By employing existing GAR methods for ensembling with our proposed method, the GAR results outperform the similar SOTA GAR methods.

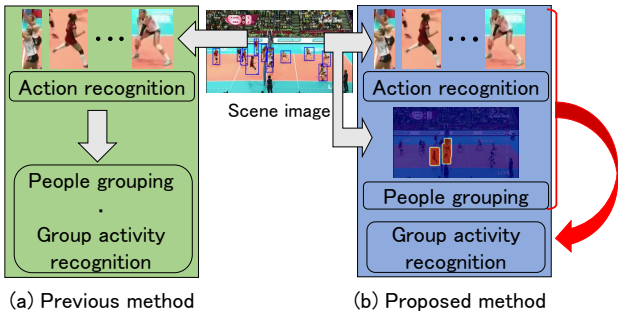

Figure 1. Comparison of GAR approaches. (a) Typical previous GAR with stepwise processes. (b) Proposed GAR method using joint learning.

2 Related Work

Individual action recognition (IAR): IAR classifies the action class of each person. Two-Stream Convnet (TSC) methods [5, 6, 7, 8] recognize individual actions using RGB and flow images. Temporal Segment Networks (TSN) [9] sparsely extracts subsequences from videos and uses each subsequence for IAR. TSN learns the entire video while reducing the computational cost. In our paper, we use TSC [6] and TSN [9] as an IAR component because of their compactness and high accuracy.

People grouping: In our paper, a "people group" is a set of people who play an important role in each scene. Previous methods use person tracking for people grouping that is based on the similarity of people's positions and directions [10]. Such methods based only on simple geometric information are effective only if people's movements can be categorized into several patterns. However, it is difficult to apply such methods to scenes where the people group changes in a complex manner (e.g., team sports). Therefore, in our paper, we use an appearance-based grouping method [11, 12], which groups people based on local and global image features in addition to the simple geometric information.

Group activity recognition (GAR): GAR recognizes an activity done jointly by multiple people. Previous methods are mainly based on graphical models such as MRF [1], AND/OR graphs [2, 3], and hierarchical models [4, 13]. A graphical model can represent difficult situations, but it is difficult to apply it to a dynamic environment where the number of people

2021 17th International Conference on Machine Vision and Applications (MVA) | 978-4-901122-20-7/21/$31.00 ©2021 10.23919/MVA51890.2021.9511390

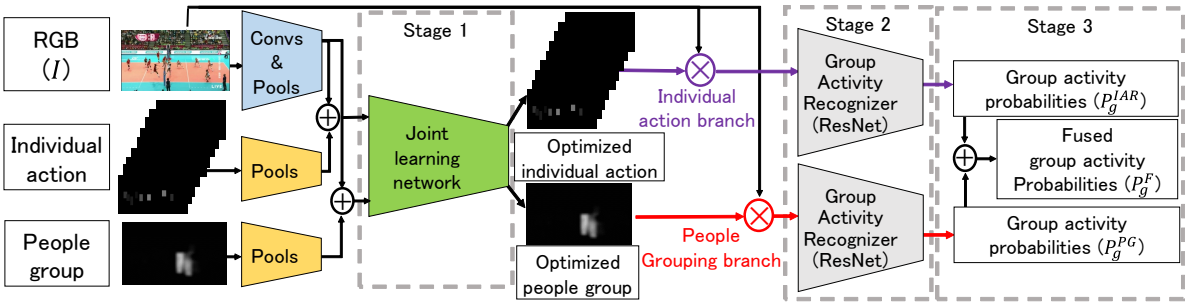

Figure 2. Our joint learning network for GAR. Three modalities (i.e., RGB image and heatmaps representing individual actions and people group on the leftmost side) are fed into the network consisting of three stages.

changes. In team-sport videos, the number of people in a field of view (FoV) is variable in general because (1) multiple players are temporarily undetectable due to occlusion, and (2) the FoV is changed largely due to camera panning, tilting, and zooming.

Joint learning: Joint learning is used for various tasks. SPFTN [14] simultaneously learns object detection and segmentation. Bagautdinov *et al.* [15] simultaneously learns people detection, IAR, and GAR in a volleyball scene. Our proposed method simultaneously learns IAR and people grouping for GAR.

3 Proposed Method

Our proposed three-stage network is shown in Fig. 2. In stage 1, individual actions and a people group obtained in advance (Sec. 3.1 and Sec. 3.2) are optimized through joint learning. In stage 2, the group activity is recognized through the individual action branch and the people grouping branch independently. In stage 3, the group activity is finalized by fusing the results of the two branches.

3.1 Individual Action Recognition

Assume that the bounding box and action class of each person are given in the dataset. A network for IAR (i.e., TSC [6] or TSN [9] in our experiments) is trained with the bounding boxes, as shown in Fig. 3 (a). This IAR network predicts the action probabilities of each person, as shown in Fig. 3 (b). With the action probabilities of all people, the heatmap of each action is obtained so that the bounding box of each person doing the corresponding action is activated. This heatmap is generated for each action, as shown in Fig. 3 (c). For example, in the heatmap of "spiking" in Fig. 3 (c), two people are detected, and the detection probability in each box is substituted in the box. With this heatmap representation, we avoid the problem of the graphical model approach that has difficulty in representing the variable number of people in a FoV.

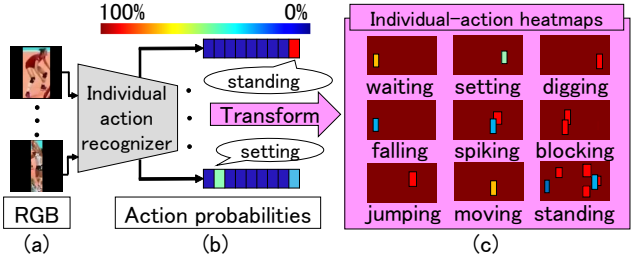

Figure 3. Heatmap representation of action probabilities. These individual-action heatmaps are fed into the network, as shown in Fig. 2.

Figure 4. (Left) People involved in the same group activity (i.e., a spiking player and two blocking players in this example) are enclosed by bounding boxes. (Right) The boxes are filled and blurred for the ground-truth annotation.

3.2 People grouping

For training the people grouping method [11], the ground-truth heatmap of each people group is annotated, as shown in the right of Fig. 4. A binary human bounding-box image is generated so that pixel values inside each person bounding-box and other pixels are set to 1 and 0, respectively. These bounding-boxes are detected by SSD [16]. Given a set of RGB images, flow images, human bounding-box images, and ground-truth heatmaps, the people grouping network, as shown in Fig. 5, is trained with the following Binary Cross Entropy loss function:

$$\sum_i \left(-G_i log(s(E_i)) - (1 - G_i) log(1 - s(E_i))\right), \qquad (1)$$

where E_i and G_i denote the i-th pixel values of the estimated and ground-truth heatmaps, respectively. $s(\cdot)$

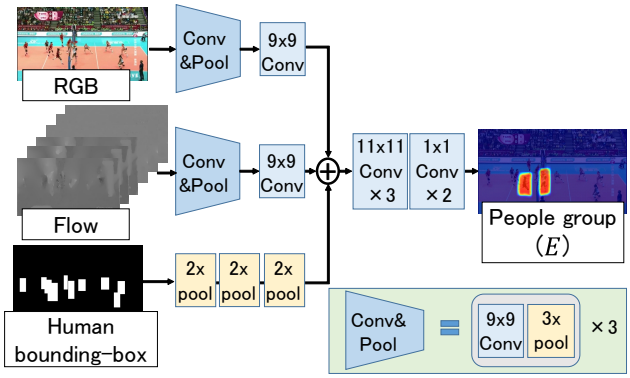

Figure 5. The architecture of people grouping network. The network estimates a people group from three modalities. (i.e, RGB, Flow, and Human bounding-box image)

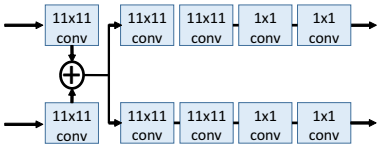

Figure 6. The architecture of joint learning network. The network optimizes individual actions and a people group simultaneously.

is the sigmoid function.

During inference, the pixel values in the estimated heatmap are binarized with a predefined threshold; 128 in our experiments. The binarized heatmap is passed to subsequent processes.

3.3 Group Activity Recognition

Our proposed three-stage network is shown in Fig. 2. In stage 1, the joint learning network optimizes individual actions (obtained in Sec. 3.1) and a people group (obtained in Sec. 3.2) through joint learning. In stage 2, the group activity is recognized through the individual action branch and the people grouping branch independently. In stage 3, the group activity is finalized by fusing the results of two branches. Our joint learning can be integrated with any IAR methods.

In stage 1, individual actions (obtained in Sec. 3.1) and a people group (obtained in Sec. 3.2) are optimized through the joint learning network, as shown in Fig. 6. Note that the heatmaps representing the individual actions and people group are normalized between 0 and 1, while the individual-action heatmaps are colored for visibility in Fig. 3. After the first layer of the network, the feature maps from individual actions and a people group are concatenated, as depicted by $\oplus$ in Fig. 6. The concatenated feature maps are fed into the second layer. Since this concatenation can be regarded as the early fusion, it is expected that their errors are mutually corrected.

In stage 2, the group activity is recognized in the individual action branch and the people grouping branch independently. We use ResNet [17] as recognition networks in these two branches. Each recognition network accepts an overlaid image (denoted by O) obtained by the following equation (depicted by $\otimes$ in Fig. 2):

$$O_i = \begin{cases} I_i \times H_i, & H_i > 0.2 \\ I_i \times 0.2, & \text{otherwise} \end{cases} \quad (2)$$

where I_i and H_i denote the i-th pixel values of the input RGB image, I, and the optimized heatmap given to each branch from the joint learning network, respectively. In O, we can emphasize the important regions in the input RGB image while keeping the global context represented in the image.

In stage 3, the final GAR result is acquired by ensembling. Specifically, this ensembling is done as the weighted sum of the GAR results (i.e., the probability values of each group activity, which are denoted by P_g^{IAR} and P_g^{PG} in Fig. 2) of the individual action and people grouping branches in the ratio of 7:3. These weights were determined empirically.

Our network consisting of stages 1, 2, and 3 is trained with the following two types of loss functions. First, the BCE loss expressed by Eq. (1) is used for each of the outputs of the joint learning network in stage 1. Second, the cross-entropy loss is computed for the final results of GAR denoted by P_g^F in stage 3. These loss functions allow us to improve GAR in stage 3 while enhancing IAR and people grouping in stage 1.

4 Experimental Results

We evaluated our proposed method with a publicly available volleyball dataset [18]. The dataset contains 4,830 sequences extracted from 55 volleyball games. These sequences are divided into 3,493 training sequences and 1,337 test sequences. A middle frame of each sequence is annotated with 9 individual action classes (i.e., waiting, setting, digging, falling, spiking, blocking, jumping, moving, and standing) and a people group, as shown in the right of Fig. 4. Each sequence is annotated with 8 group activity classes (i.e., right set, right spike, right pass, right pass, right winpoint, left set, left spike, left pass, and left winpoint).

IAR and GAR are evaluated with Multi-class Classification Accuracy (MCA) and Mean Per Class Accuracy (MPCA) [19]. People grouping is evaluated with IoU, Precision, Recall, and F-measure. In our experiments, the IoU threshold was 50%.

4.1 Individual Action Recognition

IAR results are shown in Table 1. The upper six rows show the results of TSN and TSC, each of which is executed with three different modalities (i.e., RGB, flow, and RGB+flow images). The bottom row shows

Table 1. Comparisons of IAR results (%). Our joint learning improves the IAR results.

Method	Input	MCA	MPCA
TSN	RGB	**85.1**	66.3
TSN	FLOW	82.0	55.6
TSN	RGB+FLOW	84.9	62.7
TSC	RGB	83.4	57.4
TSC	FLOW	80.3	44.2
TSC	RGB+FLOW	82.9	50.4
Joint learning (Proposed method)	RGB, FLOW	84.4	**67.2**

Table 2. Comparisons of people grouping results (%). JL means Joint Learning.

JL	Precision [%]	Recall [%]	F-measure [%]
No [11]	81.3	79.6	80.4
Yes	81.3	**80.6**	**80.9**

the best results of our joint learning network. The result is acquired when TSN using the RGB+flow images are employed as the IAR component in our proposed method.

The best MCA is acquired by TSN with RGB image without joint learning (line 1 in Table 1). On the other hand, the best MPCA is acquired by our joint learning (line 7 in Table 1). In our dataset, the number of people in each IAR class is inequality. For example, the number of "standing" accounts for about 69% of all classes. Since MCA is about 69% even if all individual classes are recognized as "standing," MPCA is more reliable than MCA in such a class imbalance situation.

4.2 People Grouping

The results of people grouping before and after our joint learning are shown in Table 2. In all three metrics, the joint learning outperforms the original results [11].

4.3 Group Activity Recognition

As shown in Table 3, we verify the performance gains by our joint learning for GAR. For GAR without the joint learning (shown in the upper three rows in the table), the original recognition results of IAR and people grouping are given directly to stage 2 by skipping stage 1. In all three cases, our joint learning outperforms the one without our joint learning.

Table 4 shows a comparison with the SOTA GAR methods. While our proposed method with stages 1, 2, and 3 outperforms CERN [20], SSU [15], and StagNet [21], PDN-ARN [22], ARG [23] and CRM [24] are better than ours.

As mentioned in Sec. 3.3, our joint learning can be integrated with any IAR methods. For improving the performance of the proposed method, the IAR result given by ARG is fed into our proposed method. Its

Table 3. GAR results (%) of our proposed method. P_g^{PG}, P_g^{IAR}, and P_g^{F} are group activity probabilities, shown on stage 3 in Fig. 2.

Method	Joint Learning	MCA	MPCA
P_g^{PG}	No	85.4	86.0
P_g^{IAR}	No	87.7	87.5
P_g^{F}	No	89.7	89.9
P_g^{PG}	Yes	85.6	86.1
P_g^{IAR}	Yes	89.4	89.1
P_g^{F}	Yes	**90.7**	**90.6**

Table 4. Comparison with other SOTA GAR methods. Our proposed method is improved by another IAR result and GAR ensembling.

Method	MCA	MPCA
Proposed method	90.7	90.6
With IAR from ARG	90.8	90.9
3-fusion	93.0	93.3
5-fusion	**93.3**	**93.6**
CERN [20] (CVPR2017)	83.3	83.6
SSU [15] (CVPR2017)	89.9	-
StagNet [21] (ECCV2018)	89.3	-
PDN-ARN [22] (MMM 2021)	92.2	-
ARG [23] (CVPR2019)	92.6	-
CRM [24] (CVPR2019)	93.0	-

results are shown in line 2 of Table 4. While its performance is better than that of the proposed method, it is still inferior to PDN-ARN, ARG and CRM. For further improvement of the proposed method, more GAR results are employed for ensembling in stage 3.

The 3-fusion method (line 3 in Table 4) fuses the GAR results of ARG in addition to those of our IAR and people grouping branches in our proposed method. The 5-fusion method (line 4 in Table 4) fuses the GAR results of the 3-fusion method and IAR and people grouping branches in the "With IAR from ARG" method (line 2 in Table 4). We can see that the 5-fusion method outperforms all SOTA GAR methods.

5 Concluding Remarks

We proposed our joint learning of IAR and people grouping for improving GAR. Our experimental results demonstrated that our joint learning optimizes individual actions and a people group, and finally improves GAR. The base method can be improved by ensembling the results of any existing GAR results.

Future work includes the improvement of components in the proposed method. In particular, people grouping is trained in a supervised manner. Since this makes it difficult to train a huge number of data, unsupervised people grouping is needed. IAR and GAR can be improved with additional cues such as estimated human poses [25, 26]. In addition, visual explanation [27] is effective for misclassification analysis.

References

[1] Zhenhua Wang, Qinfeng Shi, Chunhua Shen, and Anton van den Hengel. Bilinear programming for human activity recognition with unknown MRF graphs. In *CVPR*, 2013. 1

[2] Mohamed R. Amer, Dan Xie, Mingtian Zhao, Sinisa Todorovic, and Song Chun Zhu. Cost-sensitive top-down/bottom-up inference for multiscale activity recognition. In *ECCV*, 2012. 1

[3] Wongun Choi and Silvio Savarese. A unified framework for multi-target tracking and collective activity recognition. In *ECCV*, 2012. 1

[4] Tian Lan, Yang Wang, Weilong Yang, Stephen N. Robinovitch, and Greg Mori. Discriminative latent models for recognizing contextual group activities. *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol. 34, No. 8, pp. 1549–1562, 2012. 1

[5] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *NIPS*, 2014. 1

[6] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Val Gool. Temporal segment networks: Towards good practices for deep action recognition. In *ECCV*, 2016. 1, 2

[7] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Convolutional two-stream network fusion for video action recognition. In *CVPR*, 2016. 1

[8] Laura Sevilla-Lara, Yiyi Liao, Fatma Güney, Varun Jampani, Andreas Geiger, and Michael J. Black. On the integration of optical flow and action recognition. In *GCPR*, 2018. 1

[9] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *ECCV*, 2016. 1, 2

[10] Norimichi Ukita, Yusuke Moriguchi, and Norihiro Hagita. People re-identification across non-overlapping cameras using group features. *Comput. Vis. Image Underst.*, Vol. 144, pp. 228–236, 2016. 1

[11] Kohei Sendo and Norimichi Ukita. Heatmapping of people involved in group activities. In *MVA*, 2019. 1, 2, 4

[12] Kohei Sendo and Norimichi Ukita. Heatmapping of group people involved in the group activity. *IEICE Trans. Inf. Syst.*, Vol. 103-D, No. 6, pp. 1209–1216, 2020. 1

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

[14] Dingwen Zhang, Junwei Han, Le Yang, and Dong Xu. SPFTN: A joint learning framework for localizing and segmenting objects in weakly labeled videos. *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol. 42, No. 2, pp. 475–489, 2020. 2

[15] Timur M. 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*, 2017. 2, 4

[16] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott E. Reed, Cheng-Yang Fu, and Alexander C. Berg. SSD: single shot multibox detector. In *ECCV*, 2016. 2

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 3

[18] Mostafa S. Ibrahim, Srikanth Muralidharan, Zhiwei Deng, Arash Vahdat, and Greg Mori. A hierarchical deep temporal model for group activity recognition. In *CVPR*, 2016. 3

[19] Rui Yan, Jinhui Tang, Xiangbo Shu, Zechao Li, and Qi Tian. Participation-contributed temporal dynamic model for group activity recognition. In *ACM MM*, 2018. 3

[20] Tianmin Shu, Sinisa Todorovic, and Song-Chun Zhu. CERN: confidence-energy recurrent network for group activity recognition. In *CVPR*, 2017. 4

[21] 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*, 2018. 4

[22] Duoxuan Pei, Annan Li, and Yunhong Wang. Group activity recognition by exploiting position distribution and appearance relation. In *MMM*, 2021. 4

[23] Jianchao Wu, Limin Wang, Li Wang, Jie Guo, and Gangshan Wu. Learning actor relation graphs for group activity recognition. In *CVPR*, 2019. 4

[24] Sina Mokhtarzadeh Azar, Mina Ghadimi Atigh, Ahmad Nickabadi, and Alexandre Alahi. Convolutional relational machine for group activity recognition. In *CVPR*, 2019. 4

[25] Yuki Kawana, Norimichi Ukita, Jia-Bin Huang, and Ming-Hsuan Yang. Ensemble convolutional neural networks for pose estimation. *Comput. Vis. Image Underst.*, Vol. 169, pp. 62–74, 2018. 4

[26] Norimichi Ukita and Yusuke Uematsu. Semi- and weakly-supervised human pose estimation. *Comput. Vis. Image Underst.*, Vol. 170, pp. 67–78, 2018. 4

[27] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, 2017. 4