SUMMARY This paper proposes a method for heatmapping people who are involved in a group activity. Such people grouping is useful for understanding group activities. In prior work, people grouping is performed based on simple inflexible rules and schemes (e.g., based on proximity among people and with models representing only a constant number of people). In addition, several previous grouping methods require the results of action recognition for individual people, which may include erroneous results. On the other hand, our proposed heatmapping method can group any number of people who dynamically change their deployment. Our method can work independently of individual action recognition. A deep network for our proposed method consists of two input streams (i.e., RGB and human bounding-box images). This network outputs a heatmap representing pixelwise confidence values of the people grouping. Extensive exploration of appropriate parameters was conducted in order to optimize the input bounding-box images. As a result, we demonstrate the effectiveness of the proposed method for heatmapping people involved in group activities.

key words: action recognition, group activity, heatmap estimation, convolutional neural networks

1. Introduction

Human action recognition is one of the major topics in computer vision [1], [2]. While the actions of individuals may represent a limited amount of contextual information in an observed scene, group activities provide richer information [3]–[6]. In addition, group action recognition can support individual action recognition. In Fig. 1, for example, players enclosed by red bounding boxes are “spiking” (in the left-side court) and “blocking” (in the right-side court). However, it is not easy to recognize these individual actions only from appearance cues observed within each bounding box. In this example, these two actions (i.e., “spiking” and “blocking”) are observed synchronously in general. This property allows us to weigh the reliability of each detection, so that if any individual action included in a certain group activity is detected, the reliabilities of nearby detections whose action classes are included in this group activity. We regard such a set of individual actions as a group activity; for example, a set of “spiking” in the left court and “blocking” in the right court is defined to be “left attack” in this paper.

In complex scenes where many people behave as groups depending on their own intentions (e.g., team sports such as football and volleyball), there exists a primary group activity in general. While several people are involved in this primary group activity, other people behave depending on their own intentions. In the example of Fig. 1, three players are mainly involved in “left attack”, and other players behave depending on their own intentions. While the other players are also softly involved in this group activity (e.g., “moving” and “waiting” actions may be induced by “left attack”), their spatio-temporal synchronicities with the group activity are weaker than those of “spiking” and “blocking” that are the main actions involved in “left attack”. This paper proposes how to detect a set of people involved in a primary group activity at each frame. Furthermore, the proposed method has the following properties:

- **Heatmap representation:** While group activities are in general analyzed with graphical models in most of the previous methods [3], [6]–[12], they have difficulty in representing a dynamic increase and decrease of observed people. Since the number of people changes dynamically in general (e.g., due to a change in the field of view, moving people, and unsuccessful people detection/tracking), this paper proposes a new group representation with heatmaps.

- **Estimation independently of individual action recognition:** Our proposed method estimates the group heatmap directly from an input image, while the previous methods use the results of individual action recognition. This direct estimation allows us to complementarily employ the group information estimated by our method and the action labels recognized by another method in order to understand contextual group activities.

Our proposed grouping method works independently of individual action recognition, as illustrated in the right column of Fig. 2. On the other hand, other grouping meth-
Two types of group activity recognition schemes using people grouping. (e.g. [6]) achieves grouping and group activity recognition based on the results of preprocessed individual action recognition, as shown in the left column of Fig. 2. While our direct grouping without the individual actions is more difficult than the sequential grouping scheme with the individual actions, the direct grouping has an advantage where erroneous results of individual action recognition provide no negative impact on grouping. Furthermore, since our grouping method works framewise, it can be combined with any kind of individual recognition method independently of whether or not it works framewise [13], [14] or with temporal frames [15], [16] for group activity recognition. Compared with the conference paper of this work [17], the following new contributions are shown in this paper:

- In addition to Mean Square Error (MSE), Binary Cross Entropy (BCE) is used as an alternative for a loss function in our heatmapping network. With BCE loss, all experiments were reexecuted.
- All sequences in the dataset [4] were used, while only a limited number of sequences were used in the conference paper [17].
- Failure case analysis in our proposed method is shown in order to discuss how to improve the method more.

2. Related Work

People grouping is used in various problems such as object tracking (e.g., tracking within the field of view [18] and re-identification across the fields of view [19]).

In action recognition scenarios, group activities are analyzed mainly by graphical models such as MRF [9], and/or graphs [11], [12], and hierarchical models [7], [8]. Recently, group activities are also recognized by deep convolutional neural networks (CNNs) as with other computer vision tasks. Graphical models employ CNN features for improving discriminativity of visual cues in [3], [6]. While the graphical models allow us to acquire optimal solutions, the graphical models have difficulty in representing dynamic and flexible grouping. For example, the number of observed people is fixed in the previous methods, while (1) object detection may fail to detect some people and (2) some people may be out of the field of view due to camera panning, tilting, and zooming in an observed sequence. This limitation makes it difficult to employ these methods in realistic scenarios (e.g., recognition in broadcasting videos).

Besides the graphical models, max pooling [4] and concatenation [5] of multiple bounding boxes are used for recognizing group activities. In these approaches, however, people groups are manually defined based on simple rules (e.g., all people are observed in all frames and left- and right-side players form each group).

Our group representation using heatmaps resolves the aforementioned limitation in the number of observed people. This representation achieves more flexible people grouping driven by a primary group activity at each frame. Since the flexible representation ability in the heatmap is validated in many tasks such as object/action localization [20], saliency detection [21], human pose estimation [22], we apply its ability to the people grouping problem.

3. Heatmapping of People Involved in Group Activities

Our proposed people heatmapping employs human bounding-boxes as well as RGB images. After the human bounding-boxes are detected (Sect. 3.1), they are integrated with an RGB image in an end-to-end trainable network (Sect. 3.2) as shown in Fig. 3. The network outputs only one heatmap independently of group activities at each frame. This design is motivated by one primary goal of this grouping method, namely using the group information for group activity recognition, as shown in Fig. 2. To this end, it is reasonable to estimate one heatmap independently of group activities.

Our proposed heatmapping method is based on supervised learning so that each training image is annotated with its ground-truth heatmap. Such image heatmapping using deep networks is used in object detection [23], saliency mapping [24], and so on. In these methods, a heatmap is generated from each RGB image. The similar deep network
can estimate the heatmap representing people in a group activity from the RGB image. However, heatmapping of people involved in each group activity is more difficult than other heatmapping problems such as the aforementioned object detection and saliency mapping. This is because, for heatmapping group people, people whose appearances are similar are divided into those who are involved and not involved in a group activity.

For robust mapping, in addition to the RGB image, our proposed method employs human bounding boxes detected by an object detector. This is because 1) the deployment of players is an important cue for people grouping and 2) human detection using recent CNNs is reliable. It might be theoretically possible that a set of wide and deep layers can be used instead of the proposed network with RGB and bounding-box streams. That is, the wide and deep layers can extract both visual features and the deployment of players. However, the wide and deep layers require a large amount of training data for avoiding overfitting. Instead, in our proposed method, the bounding-box image is employed for a grouping task as a reliable common knowledge.

3.1 Human Detection and Tracking for Bounding-Box Extraction

Given a frame as an input, the bounding box of each person can be detected by a generic object detector such as SSD [25]. In our method, only person bounding-boxes are used.

If a sequence of frames is given as an input, we can improve the results of people detection using visual tracking robustly to occlusion. In this case, SSD is initially employed for people detection at each frame. The detected bounding-boxes are connected through all frames by data association [26]. We extract bounding boxes that satisfy all of the following conditions:

- Bounding boxes are tracked through all frames. While human tracking is utilized for robustly extracting human regions, the proposed method is performed at each frame.
- Bounding boxes are observed inside the court. This condition allows us to neglect umpires, coaches, audiences, and other people.

Since a set of short sequences were used in experiments shown in this paper, the latter detection-and-tracking approach was utilized.

3.2 Heatmap Estimation from RGB and Bounding-Box Images

As mentioned at the beginning of Sect. 3, the group heatmap is estimated from RGB and bounding-box images in the proposed method. The examples of RGB and bounding-box images are shown in the lefthand side in Fig. 3 that illustrates the network architecture. While similar channels such as RGB channels of an image are fed into the same convolution layer in general, different modalities are usually fed into different streams; for example, RGB and flow streams for action recognition [27]. In our problem also, the complexities of visual information observed in these two images are different; the RGB image is much complex. In order to absorb this difference, 1) these two images are fed into different streams and 2) only RGB stream consists of repetitive convolution and pooling layers before it is merged with the bounding-box image. This merged feature is fed into the final convolution layers in order to output the heatmap. Since repetitive convolution and pooling layers reduce the spatial size of the output heatmap, it is magnified to the input size by simple linear interpolation.

While we refer to a human keypoint detection using heatmaps [22] in terms of the organization of convolution and pooling layers, the problem settings in human keypoint detection [22] and ours are different so that the number of body keypoints is fixed while the number of people in group activities is changed. Furthermore, the scales of people are possibly changed in the grouping problem while each keypoint is defined as a (fixed-size) point in keypoint detection. For such flexible group heatmapping, we explore an appropriate parameter configuration, as shown in Sect. 4.

Given a set of RGB images, bounding-box images, and ground-truth heatmap images in the training stage, the network shown in Fig. 3 is trained by a loss function with Mean Squared Error (MSE) in Eq. (1) or Binary Cross Entropy (BCE) in Eq. (2):
\[
\sum_i (E_i - G_i)^2, \quad (1)
\]
\[
\sum_i (-G_i \log(f(E_i)) + (1 - G_i) \log(1 - f(E_i))), \quad (2)
\]

where \(E_i\) and \(G_i\) denote \(i\)-th pixel value in estimated and ground-truth heatmaps, respectively. \(f()\) is the sigmoid function.

Basically, BCE is employed for a classification problem, while MSE is selected for a regression problem. A continuous pixel value is regressed pixelwise in our heatmapping method. However, the original ground-truth heatmap is just a binarized image where people who are involved in a group activity are highlighted. This kind of binarized images can be represented as a result of pixelwise binary classification. Consequently, BCE is expected to be a better choice than MSE for our heatmapping problem. The effects of these two loss functions are evaluated in Sect. 4.

4. Experimental Results

4.1 Dataset

Our proposed method is evaluated with the volleyball dataset [4]. We conducted experiments with 3493 training sequences (including 296 validation sequences) and 1337 test sequences. Each sequence has ten frames. The size of each frame is \(576 \times 324\) pixels, which is shrunk from the original size in the dataset. From each frame, a region with \(500 \times 300\) pixels is randomly extracted and fed into the network. This random extraction is implemented for improving the generalization ability of the trained model. We manually annotated the bounding boxes of all players by VATIC [28].

The class of the primary group activity in each sequence and players involved in this group activity were also labeled manually. The group activity classes are “pass,” “set,” “attack,” and “winpoint.” The players involved in each group activity are defined as follows:

**Pass:** Players who are trying an underhand pass independently of whether or not they are successful.

**Set:** Player who is doing an overhand pass and those who are going to spike the ball whether they are really trying or faking.

**Attack:** Players who are spiking and blocking.

**Winpoint:** All players in the team that scores the point.

This group activity is observed right after the score.

Each of these four classes is divided into those observed in the left-side and right-side courts. In total, eight group activity classes are defined: left/right pass, left/right set, left/right attack, and left/right winpoint. One of these eight classes is given to each sequence.

Sample images, human bounding-box images, and ground-truth heatmaps are shown in Fig. 4. While the bounding boxes of all people are activated in each human bounding-box image, only people involved in each group activity are activated in the heatmap.

**Fig. 4** Sample temporal frames in each sequence. From left to right, \(t\)-th, \((t + 2)\)-th, and \((t + 4)\)-th frames are shown. In each group activity class, RGB images, bounding-box images, and ground-truth heatmaps are shown in upper, middle, and lower rows, respectively. We can see people involved in each group activity by comparing RGB images and their corresponding heatmaps. Note that each human bounding-box image is not blurred in these samples because the best grouping performance is observed with a non-blurred image among several blur parameters as shown in Table 2.
The Accuracy of the heatmapping process is evaluated by the Intersection over Union (IoU) between the estimated heatmap and its ground truth. Since IoU can be applied to the estimated heatmaps and ground-truth heatmaps, respectively. The best mean IoU is colored by red.

The IoU is indicated in percentage terms. All bounding boxes are Gaussian blurred. $\sigma$ denotes the standard deviation of the Gaussian blur. W, P, Sp, and Se denote “winpoint”, “pass”, “attack”, and “set”, respectively. The best mean IoU is colored by red.

### Mean IoUs by different thresholds for binarization of the heatmap images.

| $\sigma$ | P   | S   | A   | W   | Mean |
|----------|-----|-----|-----|-----|------|
| 0        | 31.0| 56.7| 46.3| 58.6| 45.8 |
| 10       | 27.6| 51.9| 46.8| 58.8| 43.6 |
| 20       | 25.2| 51.0| 44.3| 57.9| 41.8 |
| 30       | 25.8| 47.4| 46.6| 58.6| 41.7 |

### Mean IoUs by different parameter settings for bounding box images.

| $\sigma$ | P   | S   | A   | W   | Mean |
|----------|-----|-----|-----|-----|------|
| 0.0      | 31.0| 56.7| 46.3| 58.6| 45.8 |
| 0.1      | 27.6| 51.9| 46.8| 58.8| 43.6 |
| 0.2      | 25.2| 51.0| 44.3| 57.9| 41.8 |
| 0.3      | 25.8| 47.4| 46.6| 58.6| 41.7 |

### losses Functions for Heatmapping

4.2 Loss Functions for Heatmapping

The Accuracy of the heatmapping process is evaluated by the Intersection over Union (IoU) between the estimated heatmap and its ground truth. Since IoU can be applied to binary images, the estimated heatmap must be binarized by thresholding as illustrated in Fig. 5.

$$\text{IoU} = \frac{M_E \cap M_G}{M_E \cup M_G} \times 100 \tag{3}$$

where $M_E$ and $M_G$ denote the pixel sets of the estimated and ground-truth heatmaps, respectively.

For comparing the results of these two loss functions, estimated heatmaps were binarized so that the threshold was 100; the range of pixel value in the heatmap is between 0 and 255. The IoU accuracy is evaluated with all pixels observed in each image. The network shown in Fig. 3 was trained by the Adam optimizer with the learning rate = $10^{-4}$ and the batch size was 30.

As shown in Eqs. (1) and (2), we evaluated these two loss functions in our experiments. We evaluated the effectiveness of each loss function, as shown in Table 1 where IoUs of the two loss functions are provided for four group activities. The IoUs of BCE are greater than those of MSE in all of four group activities. Since BCE is better than MSE for our heatmapping problem as we expected, BCE is used for all experiments in what follows.

### Heatmap Thresholding for IOU Evaluation

4.3 Heatmap Thresholding for IOU Evaluation

Figure 6 shows the mean IoUs with different thresholds on the validation sequences. Based on this result with the validation sequences, the threshold was determined to be 40 for all of the following experiments with the test sequences. While the threshold was fixed to be 40 in our experiments, it can be seen that mean IoUs between a wide range of thresholds (i.e., between 20 and 70) are are similar to each other. This insensitivity of the mean IoU to the variation of the threshold is a good property of our proposed method.

Figure 6 shows the mean IoUs with different thresholds on the validation sequences. Based on this result with the validation sequences, the threshold was determined to be 40 for all of the following experiments with the test sequences. While the threshold was fixed to be 40 in our experiments, it can be seen that mean IoUs between a wide range of thresholds (i.e., between 20 and 70) are similar to each other. This insensitivity of the mean IoU to the variation of the threshold is a good property of our proposed method.

4.4 Effects of Bounding-Box Parameters

For improving the robustness to the variation in the locations and scales of the people, the bounding-box images can be blurred. In experiments shown in this section, we adjust the standard deviation of the Gaussian blur, $\sigma$, for blurring the bounding boxes of players.

Experimental results shown in Table 2 were referred to for exploring appropriate parameters in order to adjust bounding-box images for group heatmapping. This table shows the average IoUs of four group activities and their mean IoUs in different parameter settings.

In Table 2, it can be seen that $\sigma = 0$ is the best parameter though $\sigma$ does not affect the performance so much. This insensitivity of the performance to the variation of $\sigma$ is a good property of our proposed method. $\sigma = 0$ is used in all of the following comparative experiments.

### Effects of Bounding-Box Images

4.5 Effects of Bounding-Box Images

Ablation studies were conducted for validating the effectiveness of the bounding-box images. Table 3 show the IoUs acquired by the following input sources; RGB and bounding-box images (proposed), RGB images, and bounding-box images. The best performance is acquired by a set of RGB and bounding-box images. From these results, the effectiveness of the bounding-box image and its combination with the RGB image is validated.

### Comparative Experiments

The proposed method is quantitatively compared with prior
Table 3  Mean IoUs by different input sources. The results were acquired from RGB images, bounding-box images, and their combinations. The best mean IoU is colored by red.

| Input types          | P  | S  | A  | W  | Mean |
|----------------------|----|----|----|----|------|
| RGB & Bounding-box   | 31.0 | 56.7 | 46.3 | 58.6 | **45.8** |
| RGB                  | 23.2 | 41.4 | 42.0 | 49.9 | **36.8** |
| Bounding-box         | 4.8  | 10.8 | 11.4 | 13.9 | **9.4**  |

Table 4  Mean IoUs by different sources of people bounding-boxes. The results were acquired from human detection results. The best mean IoU is colored by red.

|          | P  | S  | A  | W  | Mean |
|----------|----|----|----|----|------|
| Hierarchical model [12] | 27.7 | 12.1 | 18.1 | 33.6 | **22.9** |
| Discriminative model [19] | 10.4 | 8.3  | 16.0 | 28.7 | **15.9** |
| Our method | 31.0 | 56.7 | 46.3 | 58.6 | **45.8** |

work using proximity among people [12], [19]. Note that both of [12], [19] are designed to employ temporal trajectories of people while our proposed method works at each frame. For a fair comparison, features only extracted from each frame were used in [12], [19] in our experiments. In addition, in [12], [19], only the group closest to the center of the ground-truth, \( M_G \), is regarded as \( M_E \) and used for computing the IoU, Eq. (3), while all people are divided into several groups in [12], [19].

Table 4 shows the results. Our proposed method outperforms other methods because they employ only location cues (i.e., \( x-y \) positions of people) while our method utilizes rich visual cues.

4.7 Qualitative Performance

Figure 7 shows several examples of estimated group heatmaps. Since the estimated heatmap is binarized for evaluating the IoU with its ground-truth, the binarized heatmaps are shown in the figure. For visual comparison, the ground-truth heatmaps are also shown.

While the mean IoUs of the proposed method are not sufficiently high yet as shown in Tables 2 and 3, it can be seen in Fig. 7 that the estimated heatmaps can roughly capture the locations of people involved in the primary group activity.

Typical failure cases are shown in Fig. 8.

In many failure cases of Winpoint, players of a team that scored the point get together. Therefore, if few players get together for any reason, grouping in Winpoint may fail as shown in Fig. 8 (d).

In the example of Pass shown in Fig. 8 (a), a target player who is diving for Pass (enclosed by a rectangle) cannot be detected maybe because this kind of diving action is not sufficiently included in training data.

On the other hand, this unsuccessful recognition result

---

\(^1\) Recent group activity recognition methods using CNNs and graphical models [3]–[6] cannot be compared with our proposed method because they do not provide the group information and/or individual actions are required to be recognized. On the other hand, our method explicitly extracts the group information without individual action labels. As mentioned in Introduction, we regard this property as the advantage of our method.

---

Fig. 7  Examples of group heatmaps for four group activities. In each group activity, the left and right images respectively show the estimated and ground-truth heatmaps overlaid on their corresponding image. Two rows show the heatmaps of different frames in the same sequence.
in Set is observed due to an early pass action of one of the players. In this example, this player (enclosed by a rectangle) might be considered to be doing Pass based on her pose.

In the example of Attack, shown in Fig. 8 (c), the regions of many players are activated by mistake. In our dataset, such many people are annotated only in Winpoint and Set. But in this case, these players do not get together. So, rather than Winpoint, Set may be the misrecognized activity. If such error is occurred due to the similarity of the players’ configuration between different activities, we have to concentrate more on image cues for resolving this kind of error. For example, informative body-pose cues [29]–[31] can improve individual action recognition [14]–[16].

5. Concluding Remarks

This paper proposed a method for heatmapping people who are involved in each group activity. As a novel contribution in this work, the heatmap-based group representation allows us to extract the people involved in the same group activity at each frame independently of the dynamic change in the number and deployment of the people.

Future work includes extensions using temporal frames, while our proposed method uses only one frame for people grouping at each frame. As proposed in prior work [5], [6], temporal processing using deep networks (e.g., LSTM [32] and 3DCNN [33]) is useful for people grouping in videos. Optical flows [34] also help to understand the dynamic motions of people. For such temporal processing, people tracking is crucial for inter-frame object identification. People tracking with crowded people in sport scenes should be improved by more constraints (e.g., high-order temporal smoothness [35]). Progressive improvement of the heatmap (e.g., for pose estimation [22] and for attention localization [36]) is also a promising extension. While only standard MSE loss and BCE loss are evaluated in this paper, we found unbalanced loss values among group heatmaps where the number and/or scales of group people are unbalanced. This problem can be suppressed by other loss functions such as the IoU loss [37], [38] that is applied to semantic image segmentation, which is similar to our problem (i.e., heatmapping of multiple regions). As described in Sect. 1, we assume that one primary group activity is observed at each frame. Grouping in more complex scenes/scenarios where multiple group activities are observed is also a future target. A dataset with “no group activity” class is also interesting because of the large variation of this class. The last future work is group activity recognition with people grouping.

References

[1] H. Idrees, A.R. Zamir, Y.-G. Jiang, A. Gorban, I. Laptev, R. Sukthankar, and M. Shah, “The THUMOS challenge on action recognition for videos “in the wild”,” CVIU, vol.155, pp.1–23, 2017.
[2] S. Herath, M.T. Harandi, and F. Porikli, “Going deeper into action recognition: A survey,” Image Vision Comput., vol.60, pp.4–21,
1217

[3] Z. Deng, A. Vahdat, H. Hu, and G. Mori, “Structure inference machines: Recurrent neural networks for analyzing relations in group activity recognition,” CVPR, pp.4772–4781, 2016.

[4] M.S. Ibrahim, S. Muralidharan, Z. Deng, A. Vahdat, and G. Mori, “A hierarchical deep temporal model for group activity recognition,” CVPR, pp.1971–1980, 2016.

[5] M.S. Ibrahim and G. Mori, “Hierarchical relational networks for group activity recognition and retrieval,” ECCV, vol.11207, pp.742–758, 2018.

[6] M. Qi, J. Qin, A. Li, Y. Wang, J. Luo, and L.V. Gool, “Stagnet: An attentive semantic RNN for group activity recognition,” ECCV, vol.11214, pp.104–120, 2018.

[7] T. Lan, Y. Wang, W. Yang, S.N. Robinovich, and G. Mori, “Discriminative latent models for recognizing contextual group activities,” PAMI, vol.34, no.8, pp.1549–1562, 2012.

[8] M.R. Amer, P. Lei, and S. Todorovic, “Hirf: Hierarchical random field for collective activity recognition in videos,” ECCV, vol.8694, pp.572–585, 2014.

[9] Z. Wang, Q. Shi, C. Shen, and A. van den Hengel, “Bilinear programming for human activity recognition with unknown MRF graphs,” CVPR, pp.1690–1697, 2013.

[10] W. Choi and S. Savarese, “A unified framework for multi-target tracking and collective activity recognition,” ECCV, vol.7575, pp.215–230, 2012.

[11] M.R. Amer, D. Xie, M. Zhao, S. Todorovic, and S.-C. Zhu, “Cost-sensitive top-down/bottom-up inference for multiscale activity recognition,” ECCV, vol.7575, pp.187–200, 2012.

[12] T. Shu, D. Xie, B. Rothrock, S. Todorovic, and S.-C. Zhu, “Joint inference of groups, events and human roles in aerial videos,” CVPR, pp.4576–4584, 2015.

[13] G. Guo and A. Lai, “A survey on still image based human action recognition,” Pattern Recognition, vol.47, no.10, pp.3343–3361, 2014.

[14] N. Ukita, “Pose estimation with action classification using global-and-pose features and fine-grained action-specific pose models,” IEICE Transactions, vol.E101-D, no.3, pp.758–766, 2018.

[15] B.X. Nie, C. Xiong, and S.-C. Zhu, “Joint action recognition and pose estimation from video,” CVPR, pp.1293–1301, 2015.

[16] D.C. Luvizon, D. Picard, and H. Tabia, “2d/3d pose estimation and action recognition using multitask deep learning,” CVPR, pp.5137–5146, 2018.

[17] K. Sendo and N. Ukita, “Heatmapping of people involved in group activities,” MVA, pp.1–6, 2019.

[18] W.-L. Lu, J.-A. Ting, J.J. Little, and K.P. Murphy, “Learning to track and identify players from broadcast sports videos,” PAMI, vol.35, no.7, pp.1704–1716, 2013.

[19] N. Ukita, Y. Moriguchi, and N. Hagita, “People re-identification across non-overlapping cameras using group features,” CVIU, vol.144, pp.229–236, 2016.

[20] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, “Learning deep features for discriminative localization,” CVPR, pp.2921–2929, 2016.

[21] V. Ramanshka, A. Das, J. Zhang, and K. Saenko, “Top-down visual saliency guided by captions,” CVPR, pp.3135–3144, 2017.

[22] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, “Convolutional pose machines,” CVPR, pp.4724–4732, 2016.

[23] H. Law and J. Deng, “CornerNet: Detecting objects as paired keypoints,” ECCV, pp.765–781, 2018.

[24] J. Pan, E. Sayrol, X. Giró-I-Nieto, K. McGuinness, and N.E. O’Connor, “Shallow and deep convolutional networks for saliency prediction,” CVPR, pp.598–606, 2016.

[25] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S.E. Reed, C.-Y. Fu, and A.C. Berg, “SSD: single shot multibox detector,” ECCV, vol.9905, pp.21–37, 2016.

[26] H. Pirsiavash, D. Ramanan, and C.C. Fowlkes, “Globally-optimal greedy algorithms for tracking a variable number of objects,” CVPR, pp.1201–1208, 2011.

[27] C. Feichtenhofer, A. Pinz, and A. Zisserman, “Convolutional two-stream network fusion for video action recognition,” CVPR, pp.1933–1941, 2016.

[28] C. Vondrick, D.J. Patterson, and D. Ramanan, “Efficiently scaling crowdsourced video annotation,” IJCV, vol.101, no.1, pp.184–204, 2013.

[29] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, “Realtime multi-person 2d pose estimation using part affinity fields,” CVPR, pp.1302–1310, 2017.

[30] Y. Kawana, N. Ukita, J.-B. Huang, and M.-H. Yang, “Ensemble convolutional neural networks for pose estimation,” CVIU, vol.169, pp.62–74, 2018.

[31] N. Ukita and Y. Uematsu, “Semi- and weakly-supervised human pose estimation,” CVIU, vol.170, pp.67–78, 2018.

[32] K. Greff, R.K. Srivastava, J. Koutník, B.R. Steunebrink, and J. Schmidhuber, “LSTM: A search space odyssey,” IEEE Trans. Neural Netw. Learning Syst., vol.28, no.10, pp.2222–2232, 2017.

[33] D. Tran, L.D. Bourdev, R. Fergus, L. Torresani, and M. Paluri, “Learning spatiotemporal features with 3d convolutional networks,” ICCV, pp.4489–4497, 2015.

[34] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L.V. Gool, “Temporal segment networks: Towards good practices for deep action recognition,” ECCV, vol.9912, pp.20–36, 2016.

[35] N. Ukita and A. Okada, “High-order framewise smoothness-constrained globally-optimal tracking,” CVIU, vol.153, pp.130–142, 2016.

[36] S. Chen, B. Song, J. Guo, Y. Zhang, X. Du, and M. Guizani, “FPAN: fine-grained and progressive attention localization network for data retrieval,” Computer Networks, vol.143, pp.98–111, 2018.

[37] M. Berman, A.R. Triki, and M.B. Blaschko, “The lovazs-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks,” CVPR, pp.4413–4421, 2018.

[38] J. Yu, Y. Jiang, Z. Wang, Z. Cao, and T.S. Huang, “Unibbox: An advanced object detection network,” ACM MM, pp.516–520, 2016.

Kohei Sendo is a master course student of Toyota Technological Institute, Japan. His research theme is human action recognition and its extension to group activity recognition.

Normichi Ukita is a professor at the graduate school of engineering, Toyota Technological Institute, Japan (TTI-J). He received the Ph.D. degree in Informatics from Kyoto University, Japan, in 2001. After working for five years as an assistant professor at NAIST, he became an associate professor in 2007 and moved to TTI-J in 2016. He was a research scientist of PRESTO, JST during 2002–2006. He was a visiting research scientist at Carnegie Mellon University during 2007–2009. His main research interests are multi-object tracking and human pose and activity estimation.