16th International Conference on Machine Vision Applications (MVA)
National Olympics Memorial Youth Center, Tokyo, Japan, May 27-31, 2019.

Heatmapping of People Involved in Group Activities

Kohei Sendo Norimichi Ukita
Toyota Technological Institute

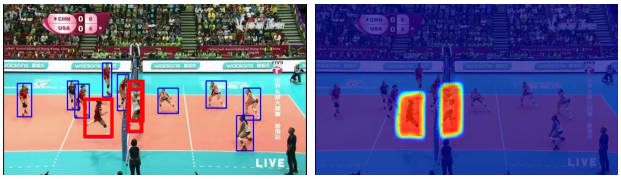
Abstract

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). Our proposed heatmapping method can group any number of people who dynamically change their deployment. A deep network for this 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.

1 Introduction

Human action recognition is one of the major topics in computer vision. While the actions of individuals may represent a limited amount of contextual information in an observed scene, group activities provide richer information [4, 8, 7, 16]. In addition, group activity recognition can support individual action recognition. In Figure 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 for one of the two actions by grouping it with a detection of the other action. 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 spike" in this paper.

In group activity scenarios such as team sports (e.g., 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 Figure 1, three players are mainly involved in "left spike", 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 spike"), their spatio-temporal synchronicities with the group activity are weaker than those of "spiking" and "blocking" that are the main actions involved in "left spike".



Input image overlaid by people bounding-boxes | Result overlaid on the image

Figure 1. Heatmapping of people in a group activity. The regions of people involved in the same group activity are extracted as a heatmap. In the heatmap, the confidence of this region extraction is given pixelwise.

This paper proposes how to detect a set of people involved in a primary group activity at each frame. 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 [9, 1, 24, 2, 18, 4, 16], 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 complementary employ the group information estimated by our method and the action labels recognized by another method in order to understand contextual group activities.

2 Related Work

People grouping is used in various problems such as object tracking (e.g., tracking within the field of view [12] and re-identification across the fields of view [20]).

© 2019 MVA Organization

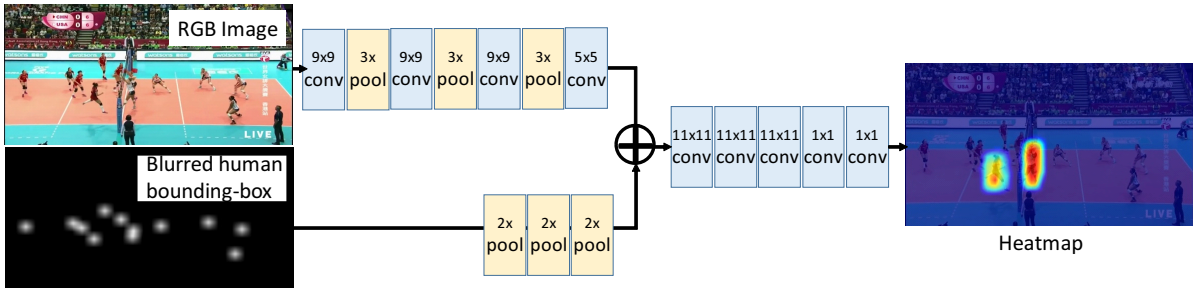


Figure 2. Proposed network for heatmapping a group. While the output of this network is a heatmap, its is overlaid on the input RGB image for visualization. The RGB and heatmap streams output 32 and 1 channels, respectively.

In action recognition, group activities are analyzed mainly by graphical models such as MRF [24], and-or graphs [2, 18], hierarchical models [9, 1]. 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 [4, 16]. 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.

Besides the graphical models, max pooling [8] and concatenation [7] 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 proposed 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 [26], saliency detection [17], human pose estimation [25], we apply its ability to the people grouping problem.

3 Heatmapping of People Involved in 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 [10], saliency mapping [14], 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 [10, 14]. 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. Note that, even if people are not successfully detected, our proposed method is designed to work with an RGB image.

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 [11]. In the proposed 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 our proposed method, given a sequence of frames, SSD is initially employed for people detection at each frame. The detected bounding-boxes are connected through all frames by data association [15]. 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 Section 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 left-hand side in Figure 2 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 [5]. 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 [25] in terms of the organization of convolution and pooling layers, the problem settings in human keypoint detection [25] 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 Section 4.

Given a set of RGB images, bounding-box images, and ground-truth heatmap images in the training stage, the network shown in Figure 2 is trained by the MSE loss function:

$$\sum_i \left(H_{E,i} - H_{G,i}\right)^2, \tag{1}$$

where $H_{E,i}$ and $H_{G,i}$ denote i-th pixel value in estimated and ground-truth heatmaps, respectively.

4 Experimental Results

Our proposed method is evaluated with the publicly-available volleyball dataset [8]. In our experiments, we conducted experiments with 347 training sequences, 23 validation sequences, and 150 test sequences. Each sequence has six frames. The size of each frame is 576 × 324 pixels, which is shrinked from the original size in the dataset. From each frame with 576 × 324 pixels, a region with 500 × 300 pixels is randomly extracted and fed into the network. This random extraction is implemented for improving the generalization ability of

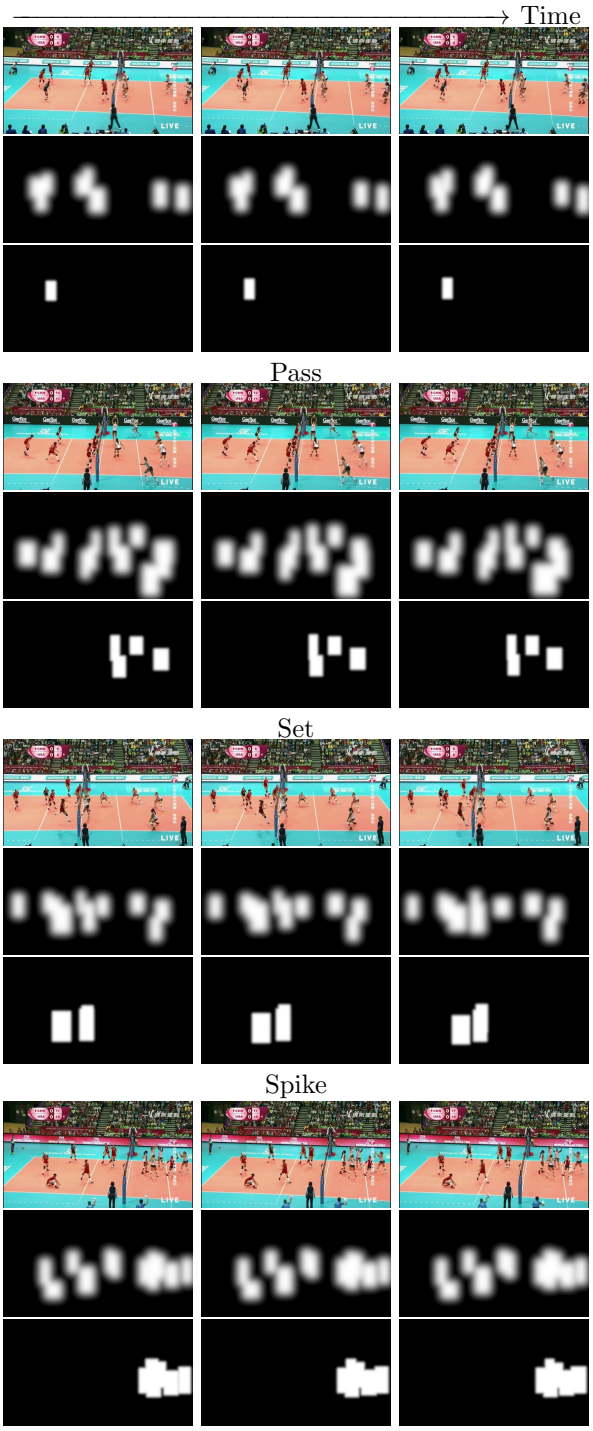


Figure 3. Sample temporal frames in each sequence. From left to right, t-th, $(t+2)$-th, and $(t+4)$-th frames are shown. In each group activity class, RGB images, bounding-box images, and ground-truth heatmaps are shown in upper, middle, and lower rows, respectively. We can see people involved in each group activity by comparing RGB images and their corresponding heatmaps.

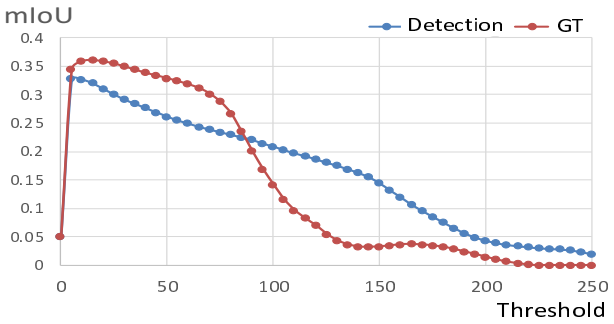


Figure 4. Mean IoUs by different thresholds for binarization of the heatmap images.

the trained model. We manually annotated the bounding boxes of all players by VATIC [22]. 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," "spike," 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.

Spike: Players who are spiking and blocking.

Winpoint: All players in the team that scores the point. This group activity is observed for a few seconds 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. One of these eight classes is given to each sequence.

Sample images, human bounding-box images, and ground-truth heatmaps are shown in Figure 3. 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. For improving the robustness to the variation in the locations and scales of the people, the bounding-box images are resized and blurred, as shown in Figure 3. The parameters of these resizing and blurring are explored in experiments shown in Section 4.2.

4.1 Heatmap Thresholding for IoU Evaluation

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.

$$\text{IoU} = \frac{\boldsymbol{M}_E \cap \boldsymbol{M}_G}{\boldsymbol{M}_E \cup \boldsymbol{M}_G} \tag{2}$$

where $\boldsymbol{M}_E$ and $\boldsymbol{M}_G$ denote the pixel sets of the estimated and ground-truth heatmaps, respectively. The IoU accuracy is evaluated with all pixels observed in each image in this paper. Note that we can also evaluate the IoU in each person bounding-box in order to validate whether or not each person is involved in the group activity.

Figure 4 shows the mean IoUs with different thresholds on the validation sequences. The mean IoUs were computed from two type of heatmaps; those estimated from ground-truth bounding-box images and those estimated from detection bounding-boxes. Based on these results with the validation sequences, the threshold was determined to be 10, which is nearly the average of the best thresholds in the two types of heatmaps, for all of the following experiments with the test sequences.

The network shown in Figure 2 was trained by the Adam optimizer with the learning rate $= 10^{-4}$ and the batch size was 8.

Table 1. Mean IoUs by different parameters for bounding box images. The IoU is indicated in percentage terms. All bounding boxes are Gaussian blurred. σ denotes the standard deviation of the Gaussian blur. W, P, Sp, and Se denote "winpoint", "pass", "spike", and "set", respectively. The best mean IoU is colored by red.

Gauss	W	P	Se	Sp	Mean
0	43.2	20.9	48.0	41.0	34.6
10	45.3	22.1	50.8	43.7	36.6
20	42.6	21.6	50.7	43.9	35.7
30	39.2	20.0	47.4	43.3	33.5

4.2 Effects of Bounding-box Parameters

Experimental results shown in Table 1 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 this paper, we adjust the standard deviation of the Gaussian blur, σ, for the bounding boxes of players.

In Table 1, it can be seen that $\sigma = 10$ is the best parameter though σ does not affect the performance so much. $\sigma = 10$ is used in all of the following comparative experiments. The bounding-box in the ground-truth heatmap is also blurred but its blur is only $\sigma = 3$. This is because, in most similar problems such as object detection, the bounding-box in the ground-truth image is not blurred.

Table 2. Mean IoUs by different sources of people bounding-boxes. The results were acquired from human detection results and their ground-truth.

Input types	W	P	Sp	Se	Mean
Detection boxes	45	22	51	44	37
Ground-truth boxes	40	21	41	50	34

Table 3. Mean IoUs by different sources of people bounding-boxes. The results were acquired from human detection results.

	W	P	Sp	Se	Mean
Hierarchical model [18]	33	27	18	12	23
Discriminative model [20]	28	10	16	8	16
Our method	40	21	41	50	34

4.3 Influence of People Detection Performance

This section shows the results of experiments with ground-truth human bounding-boxes instead of automatically-detected bounding-boxes. Both experiments were conducted with the same Gaussian blur parameter, $\sigma = 10$. Surprisingly, the mean IoU is decreased with the ground-truth bounding boxes; 34% by the detection results vs 37% by the ground-truth bounding boxes. This result validates that the proposed method is robust to failure in object detection.

4.4 Comparative Experiments

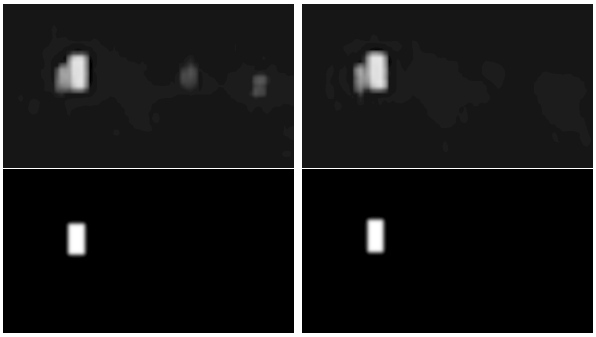
The proposed method is quantitatively compared with prior work using proximity among people [18, 20][1]. Note that both of [18, 20] are designed to employ temporal trajectories of people while our proposed method works at each frame. For a fair comparison, features only extracted from each frame were used in [18, 20] in our experiments. In addition, in [18, 20], only the group closest to the center of the ground-truth, $\boldsymbol{M}_G$, is regarded as $\boldsymbol{M}_E$ and used for computing the IoU, Eq. (2), while all people are divided into several groups in [18, 20].

Table 3 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.

4.5 Qualitative Performance

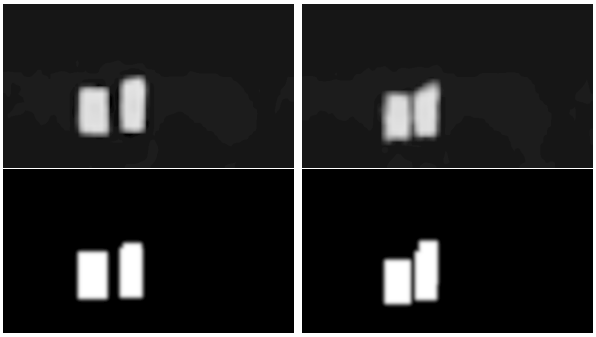
Figure 5 shows several examples of estimated group heatmaps. Since the estimated heatmap is binarized

[1]Recent group activity recognition methods using CNNs and graphical models [4, 8, 7, 16] 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.

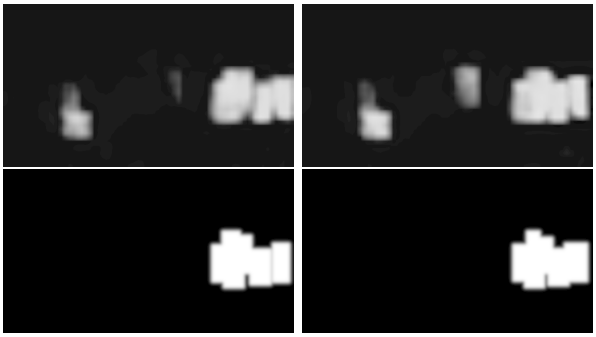


Pass (Upper: Results, Lower: Ground-truth)

Set (Upper: Results, Lower: Ground-truth)



Spike (Upper: Results, Lower: Ground-truth)



Winpoint (Upper: Results, Lower: Ground-truth)

Figure 5. Examples of group heatmaps for four group activities. In each group activity, the upper and lower images respectively show the estimated and ground-truth heatmaps overlaid on their corresponding image. Two columns show the heatmaps of different frames in the same sequence.

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 1 and 2, it can be seen in Figure 5 that the estimated heatmaps can roughly capture the locations of people involved in the primary group activity. Our future work will investigate how this group representation using the heatmap can support group activity recognition.

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 the method proposed in this paper uses only one frame for people grouping at that frame. As proposed in prior work [7, 16], temporal data processing using deep networks (e.g., LSTM [6] and 3DCNN [19]) are useful for people grouping in videos. Optical flows [23] also help to understand the dynamic motions of people in videos. For such temporal processing, people tracking is crucial for interframe object identification. People tracking with crowded people in sports scenes should be improved by more constraints (e.g., high-order temporal smoothness [21]). Progressive improvement of the heatmap (e.g., for pose estimation [25] and for attention localization [3]) is also a promising extension of our proposed method. As mentioned in Section 1, a prospective application of our proposed grouping method is group activity recognition with individual action recognition. Since our heatmapping method works framewise, individual action recognition is also expected to work framewise (e.g., [13]).

References

[1] M. R. Amer, P. Lei, and S. Todorovic. Hirf: Hierarchical random field for collective activity recognition in videos. In *ECCV*, 2014. 1, 2

[2] M. R. Amer, D. Xie, M. Zhao, S. Todorovic, and S. C. Zhu. Cost-sensitive top-down/bottom-up inference for multiscale activity recognition. In *ECCV*, 2012. 1, 2

[3] 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*, 143:98–111, 2018. 6

[4] Z. Deng, A. Vahdat, H. Hu, and G. Mori. Structure inference machines: Recurrent neural networks for analyzing relations in group activity recognition. In *CVPR*, 2016. 1, 2, 5

[5] C. Feichtenhofer, A. Pinz, and A. Zisserman. Convolutional two-stream network fusion for video action recognition. In *CVPR*, 2016. 3

[6] K. Greff, R. K. Srivastava, J. Koutník, B. R. Steunebrink, and J. Schmidhuber. LSTM: A search space odyssey. *IEEE Trans. Neural Netw. Learning Syst.*, 28(10):2222–2232, 2017. 6

[7] M. S. Ibrahim and G. Mori. Hierarchical relational networks for group activity recognition and retrieval. In *ECCV*, 2018. 1, 2, 5, 6

[8] M. S. Ibrahim, S. Muralidharan, Z. Deng, A. Vahdat, and G. Mori. A hierarchical deep temporal model for group activity recognition. In *CVPR*, 2016. 1, 2, 3, 5

[9] T. Lan, Y. Wang, W. Yang, S. N. Robinovitch, and G. Mori. Discriminative latent models for recognizing contextual group activities. *PAMI*, 34(8):1549–1562, 2012. 1, 2

[10] H. Law and J. Deng. Cornernet: Detecting objects as paired keypoints. In *ECCV*, 2018. 2

[11] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. E. Reed, C. Fu, and A. C. Berg. SSD: single shot multibox detector. In *ECCV*, 2016. 2

[12] W. Lu, J. Ting, J. J. Little, and K. P. Murphy. Learning to track and identify players from broadcast sports videos. *PAMI*, 35(7):1704–1716, 2013. 1

[13] K. Morimoto, Y. Matsuyama, and N. Ukita. Continuous action recognition by action-specific motion models. In *MVA*, 2013. 6

[14] J. Pan, E. Sayrol, X. Giró i Nieto, K. McGuinness, and N. E. O'Connor. Shallow and deep convolutional networks for saliency prediction. In *CVPR*, 2016. 2

[15] H. Pirsiavash, D. Ramanan, and C. C. Fowlkes. Globally-optimal greedy algorithms for tracking a variable number of objects. In *CVPR*, 2011. 2

[16] M. Qi, J. Qin, A. Li, Y. Wang, J. Luo, and L. V. Gool. stagnet: An attentive semantic RNN for group activity recognition. In *ECCV*, 2018. 1, 2, 5, 6

[17] V. Ramanishka, A. Das, J. Zhang, and K. Saenko. Top-down visual saliency guided by captions. In *CVPR*, 2017. 2

[18] T. Shu, D. Xie, B. Rothrock, S. Todorovic, and S. Zhu. Joint inference of groups, events and human roles in aerial videos. In *CVPR*, 2015. 1, 2, 5

[19] D. Tran, L. D. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*, 2015. 6

[20] N. Ukita, Y. Moriguchi, and N. Hagita. People re-identification across non-overlapping cameras using group features. *CVIU*, 144:228–236, 2016. 1, 5

[21] N. Ukita and A. Okada. High-order framewise smoothness-constrained globally-optimal tracking. *CVIU*, 153:130–142, 2016. 6

[22] C. Vondrick, D. J. Patterson, and D. Ramanan. Efficiently scaling up crowdsourced video annotation. *IJCV*, 101(1):184–204, 2013. 4

[23] 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. In *ECCV*, 2016. 6

[24] Z. Wang, Q. Shi, C. Shen, and A. van den Hengel. Bilinear programming for human activity recognition with unknown MRF graphs. In *CVPR*, 2013. 1, 2

[25] S. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. Convolutional pose machines. In *CVPR*, 2016. 2, 3, 6

[26] B. Zhou, A. Khosla, À. Lapedriza, A. Oliva, and A. Torralba. Learning deep features for discriminative localization. In *CVPR*, 2016. 2