小規模スポーツイベントにおける観客の盛り上がり分析手法

阿部和広，中村ちから，大坪洋介，小池哲也，横矢直人

Spectator Excitement Analysis Method in Small-scale Sports Events[†]

Kazuhiro ABE, Chikara NAKAMURA, Yosuke OTSUBO, Tetsuya KOIKE and Naoto YOKOYA

観客の盛り上がりを検知することによって自動ハイライト生成や自動映像編集などの様々な応用が可能であり，観客分析が幅広く研究されている．観客分析の手法として，全体論的手法とオブジェクトベース手法がある．先行研究の多くは全体論的手法を用いているが，大規模な試合と比べて観客の数が少ない小規模な試合では有効ではない．そこで，本研究では，オブジェクトベース手法を用いた小規模な試合における観客の盛り上がり検知手法を提案する．本手法の有効性を検証するために，観客と選手を撮影したデータセットを構築した．実験を行った結果，全体論的手法のベースラインと比べて性能が良く，観客個人の盛り上がり検知も可能になった．また，検知結果から盛り上がりシーン分析を行った結果，盛り上がりスコアが高いシーンが高得点ゴールシーンに対応していることが分かった．

The detection of the excitement of spectators in sports is useful for various applications, such as automatic highlight generation and automatic video editing. Therefore, spectator analysis has been widely studied. Two main approaches used for this include holistic and object-based approaches. Holistic approaches have been applied in most previous studies; however, they are not applicable to small-scale games, where the number of spectators is fewer compared to those of large-scale games. We herein propose a method for detecting the excitement state of spectators in small-scale games using an object-based approach. To evaluate our method, we build our own datasets comprising both spectator and player videos. Experimental results show that our method outperforms a holistic baseline method and allows the excitement detection of individual spectators. Moreover, we discovered that scenes with higher excitement scores correspond to high-score-goal scenes through the analysis of scenes pertaining to excitement using our method.

Key words スポーツ映像分析，大衆行動分析，盛り上がり検知，行動認識
sports video analysis, crowd behavior analysis, excitement detection, action recognition

1 Introduction

Understanding and analyzing crowd dynamics is important in various fields, such as surveillance, advertising, determining movie ratings, and automatic video editing. Even in sports, the reactions and motions (i.e., excitement) of spectators can be utilized to extract information regarding games because they are significantly related to the impressiveness of the sports events. In particular, the excitement state is a useful parameter as it enables one to measure the appeal of gaming events. Information regarding spectator excitement has been used for highlight generation[1)~3)] and automatic video editing[4)].

Spectator analysis has been well studied for crowd analysis. Crowd analysis has attracted attention in past decades in the field of computer vision. It has been studied in the context of crowd behavior analysis[5)], crowd density estimation[6)], and crowd motion detection[7)]. Typically, two different approaches are used in crowd analysis: holistic and object-based ones[8)]. Holistic approaches address crowds themselves rather than the details of each individual. Meanwhile, object-based approaches focus on the behavior of individuals rather than that of crowds. Both approaches can be applied for spectator analysis; however, most previous studies relied on holistic approaches[2)9)]. This is because these studies focused on large and major games, where the number of spectators is high.

In this study, we performed an excitement detection of

[†] This article contains the review of the paper[24)]

spectators for small-scale games, where the number of spectators was fewer compared with those of large-scale games. The following factors were considered in this study: 1) the density of spectators was sparse, 2) the accommodation of facilities was small, and 3) the positions of cameras were restricted. Conventional holistic approaches, although applicable to the case of dense spectators, fail in small-scale settings because of the problems above; this is because it is difficult to record videos that include only spectators due to factors 2) and 3). Hence, we propose a novel method for spectator excitement detection, which is based on an object-based approach. Our method comprises three aspects: a) upper body detection based on face detection, b) spectator classification, and c) scoring for excitement based on convolutional neural network (CNN) architectures.

As mentioned, distinguishing between spectators and other people (e.g., players or referees) is necessary, which corresponds to b). To achieve this, we performed a) as a preprocessing step. For c), we input an optical flow of spectators into a two-stream CNN[10] and defined the excitement score based on the features of the neural network.

Our contributions are summarized as follows:

- We proposed an approach to detect the state of spectator excitement for the case of sparse spectators.
- We acquired the video datasets of both spectators and players in 3 × 3 basketball games.
- We applied our method to our datasets and evaluated the performance.

This paper is organized as follows. In Section 2, we explain the details of our datasets and methods to annotate the states of spectator excitement. In Section 3, we describe our method in detail; in Section 4, results including performance evaluations are provided. Finally, we summarize our study in Section 5.

2 Dataset

(1) Dataset preparation

Several datasets are available for crowd analysis[11~13]. However, these datasets are primarily designed for public crowd analysis. In the study of spectator analysis in sports, the motion patterns of spectators are significantly related to player actions and game events. Therefore, to analyze crowd excitement, datasets that include game information corresponding to crowd motions are desirable.

Furthermore, publicly available datasets have been presented to analyze sports such as soccer[14], volleyball[15], and ice hockey[9]. In particular, the S-Hock dataset[9] is a unique dataset that captures both players and spectators simultaneously and contains dense annotations of each spectator. While the S-Hock dataset is valuable for our study, it primarily focuses on dense crowds in large games and excludes detailed appearances, which we aim to capture.

To evaluate our approach, we built our own dataset comprising videos recorded during a 3 × 3 basketball tournament organized by Alborada in Tsukuba city, Japan. This tournament included 12 games and lasted approximately 154 min.

We set up three types of cameras: spectator, field, and overlooking cameras. Six spectator cameras were used to record the motions of spectators from their seats. In addition, we set up five field cameras to record the actions of the players. Additionally, we set up four overlooking cameras to capture the overlooking view of the game. In each game, all cameras were temporally synchronized. The overall camera configuration is shown in Fig. 1. We used 4K resolution cameras to obtain fine-grained appearance information of spectators. Some example frames are shown in Fig. 2.

(2) Annotation

We annotated the moment spectator excitement detected during the game. It was difficult to determine whether an individual was excited/not excited only by watching spectator videos because they exhibited various behaviors and reactions, which varied by person. Hence, we conducted frame-level labeling based on events occurring in the field.

We assumed that the spectators were generally excited immediately after a goal was achieved, and that the excitement continued for a few seconds. After localizing the moment of the goal achievement as the time when the ball was shot into the target, we identified the excitement duration. Excitement was defined by spectator actions, such as

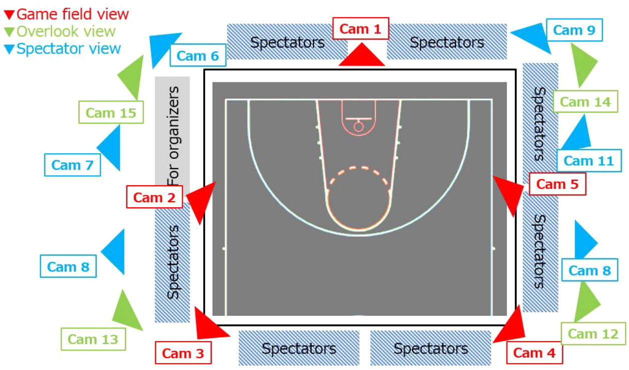
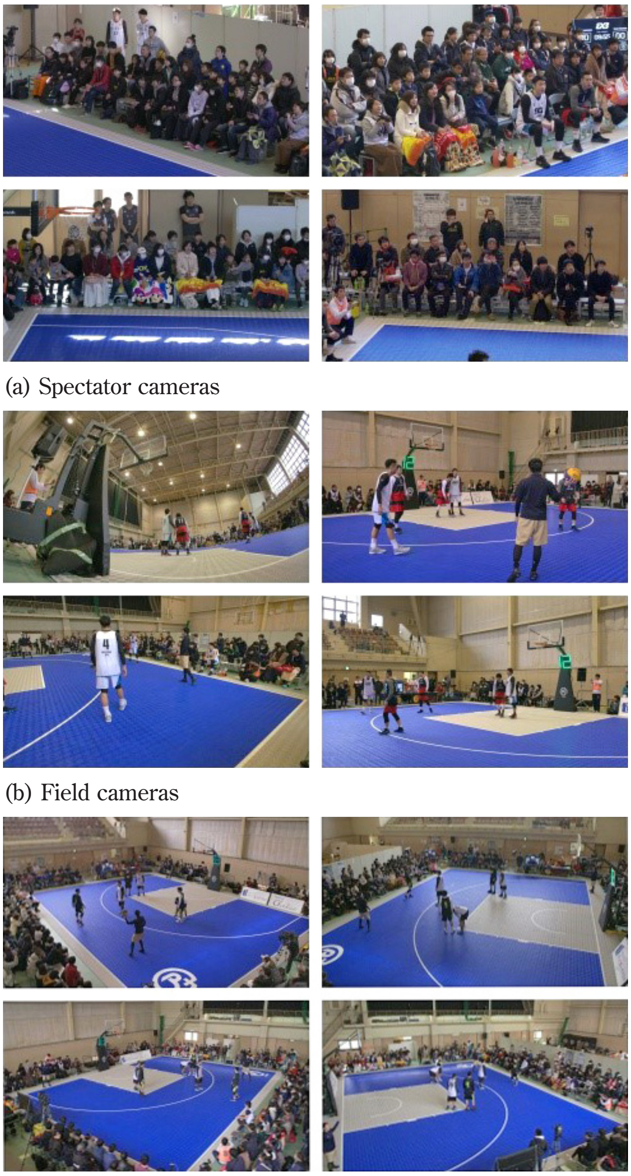


Fig. 1 Camera configuration. Red triangles indicate the field cameras that capture players. Green triangles indicate overlooking cameras that cover the overlooking view of the game. Blue triangles indicate spectator cameras that cover spectator seats.



(a) Spectator cameras

(b) Field cameras

(c) Overlooking cameras

Fig. 2 Example frame of our dataset. We obtained 4K resolution videos; however, these example images are resized to reduce image size. Videos are captured by Jiro Akiba/Getty Images.

clapping and arm raising.

3 Method

We propose an excitement detection algorithm based on an object-based approach, illustrated in Fig. 3. To capture individual appearance features, our method starts by detecting and tracking as a preprocessing step. After each person was tracked, we used a motion CNN with a two-stream architecture[10] to extract deep features; output features were generated through a trainable fully connected neural network. Subsequently, discriminative motion features of each person were aggregated to form the final frame-level score. Our motion CNN architecture can be end-to-end trained without information regarding the excitement of each person.

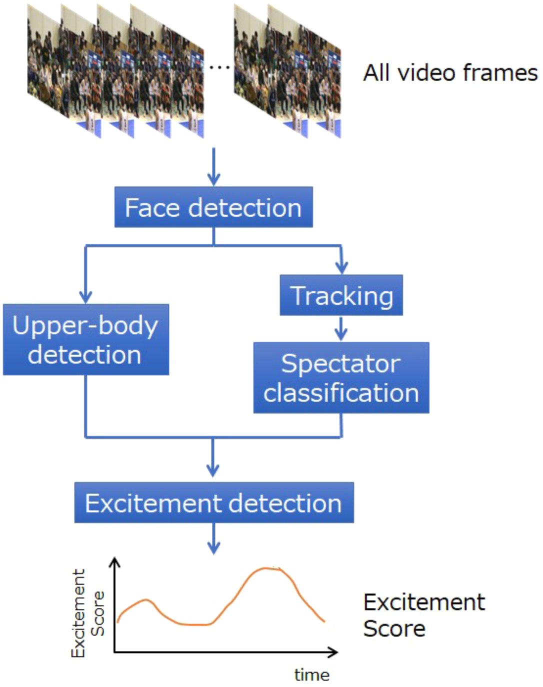


Fig. 3 Overview of our object-based excitement detection algorithm.

(1) Upper body detection based on face detection

We analyzed the features of each person based on face detection. Human detection is extremely difficult because the bodies of spectators are generally occluded and overlapped. By contrast, their faces are clearly visible and have no overlaps. Hence, it is reasonable to extract information from faces.

In general, detecting small faces is challenging. We used Single Shot Scale-invariant Face Detector (S3FD)[16] pretrained on WIDER FACE[17] as a face detector. S3FD is based on an anchor matching strategy where small faces can be detected with high accuracy. Furthermore, S3FD offers reliability in face detection. To remove misdetections, we accepted bounding boxes with reliabilities higher than 0.5.

After detecting faces, we determined the bounding box of each spectator in the following manner. Assuming the ratio of the face to the upper body regions is almost the same for each person, we determined the upper body bounding box using the following linear relationships:

$$b^B_{x1} = b^F_{x1} + a_{x1} * w, \quad (1)$$

$$b^B_{y1} = b^F_{y1} + a_{y1} * h, \quad (2)$$

$$b^B_{x2} = b^F_{x2} + a_{x2} * w, \quad (3)$$

$$b^B_{y2} = b^F_{y2} + a_{y2} * h, \quad (4)$$

where (b_{x1}, b_{y1}), (b_{x2}, b_{y2}) are the x, y coordinates of the upper-left and lower-right bounding boxes, respectively;

superscript B and F indicate the upper body and face, respectively; w and h denote the width and height of the bounding box of the face, respectively; a_{x1}, a_{x2}, a_{y1}, and a_{y2} are fitting coefficients set as $a_{x1} = -2$, $a_{y2} = -0.5$, $a_{x2} = 2$, and $a_{y2} = 4$ from empirical observation.

(2) Spectator classification

Although our target was the audience sitting on the seats, they were sometimes occluded by the players because the seats were set up close to the basketball field. In addition, some people were standing behind the seats to watch the game. Therefore, we had to distinguish the audience sitting on the seats from other people in the videos.

We classified the persons in the video into two groups: the static group and the dynamic (moving) group; the people of the latter group were excluded from spectator analysis.

First, we tracked all the persons in the videos throughout the game by the SORT tracker[18]. We then obtained the trajectories of the coordinates for each person. We classified them into two groups, in which we assigned a person into the static group if his/her coordinates of the center of the bounding box had changed within a fixed threshold; otherwise, they were assigned to the other group. We fixed the threshold parameter as 800 pixels for our 4K videos (3840 × 2160 pixels). Fig. 4 shows the qualitative result.

(3) Excitement detection

To detect the excitement state, we introduced a neural network architecture based on the temporal stream of a two-stream CNN[10]. Fig. 5 illustrates the overview of the architecture. First, following the procedures of a two-stream CNN, we prepared optical flow images containing two channels: a horizontal and a vertical component.

For a series of optical flow images, the optical flow images of 10 consecutive frames corresponding to the position of the detected upper body bounding boxes were stacked and treated as a patch. In all consecutive frames, we selected n_p patches randomly and fed them as inputs to the motion CNN. The outputs of the motion CNN were then pooled to aggregate individual scores and form the final score. We investigated the average pooling and max-pooling for person aggregation. The cross-entropy loss was used to learn from the label (excited/not excited).

We used 101-layer residual network (ResNet-101) as the backbone of the two-stream CNN architecture[19]. To avoid overfitting and leverage general motion features acquired from a large dataset, we used a model[20] pretrained for action recognition on the UCF-101 dataset[*1, 21] and fine-tuned it. Because the feature to be captured was motion, we did not use the spatial stream and only used the temporal stream in the two-stream network. For fine tuning, fully connected layers of ResNet-101 were replaced with two fully connected layers having 10 and 2 output units; the output was the softmax score.

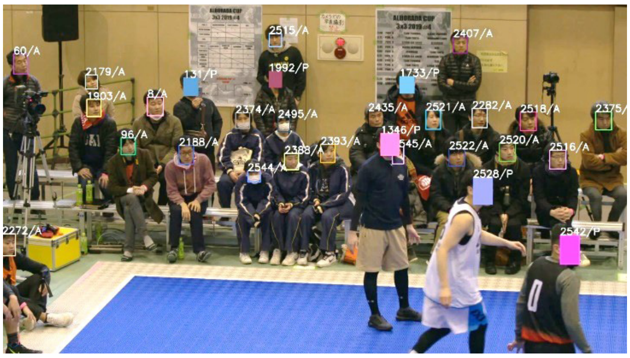


Fig. 4 Qualitative result of spectator classification. Bounding boxes of non-spectators are filled out. Original videos are captured by Jiro Akiba/Getty Images.

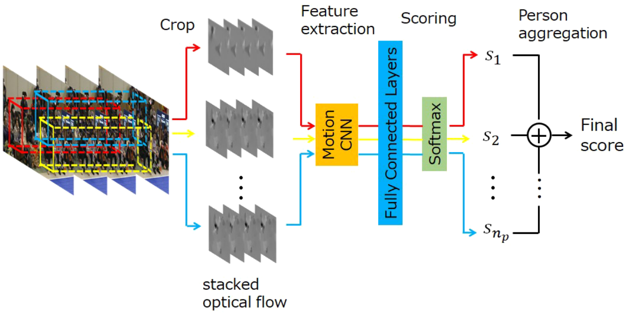


Fig. 5 Architecture of our module for excitement detection.

The optical flow was computed using the OpenCV implementation of the TV-L1 algorithm[22]. Following the implementation[20], the extracted optical flow was clipped to the range $[-20, 20]$, rescaled to the range $[0, 255]$ by linear transformation and compressed to JPEG.

In this study, we used only videos of the final and semifinal games because the spectators exhibited excitement clearly. Generally, whether one exhibits excitement depends on the team support. Therefore, we used the excitement labels from only the home team. Among 116,628 training frames, only 5,376 frames had positive labels (i.e., "excited"). For each set of 10 consecutive frames, we simply used the label of the first frame. To learn from biased data, we sampled positive and negative frames equally. We treated a set of 10 consecutive frames including n_p patches

[*1] Standard action recognition benchmark dataset.

as a batch and set the batch size to 2. We trained 20 epochs and used the Adam optimizer[23] with a learning rate of 0.0001, $\beta_1 = 0.9$, and $\beta_2 = 0.999$.

4 Evaluation

(1) Baseline method

To compare our methods with a conventional holistic method, we considered a baseline method based on a holistic approach, similar to that in[2]. This baseline method does not detect or track individuals; it is different from our method only in the cropping stage, as shown in Fig. 5. This baseline crops a fixed rectangular area (480 × 540 pixels) of the frame to form a patch and processes it as shown in Fig. 5 to obtain the final score. This patch contains one or two persons. We randomly sampled eight patches from the entire rectangular area by a sliding window with a 240 × 270 pixel (half the size of the patch) overlapped area.

(2) Comparison with baseline method

We evaluated our method on a test set corresponding to the second half of each game. Because our data were highly biased, we sampled the same number of frames from the "excited" and "not excited" classes in the following evaluations. Table 1 shows the recall, precision, and average precision of the baseline method and our proposed method. We considered different pooling methods: average pooling and max pooling. Moreover, we considered two cases of the following number of patches: n_p = 8 and n_p =16. In both cases, the training was conducted only for n_p =8. For all cases, the recall–precision curves are shown in Fig. 6.

As shown in Table 1, our method clearly outperformed the baseline method. For different types of pooling of our methods for person aggregation, the max pooling method achieved a better average precision score; however, the performance gap was relatively small. In both pooling methods, increasing n_p improved the performance, which was expected considering that more spectator information was obtained.

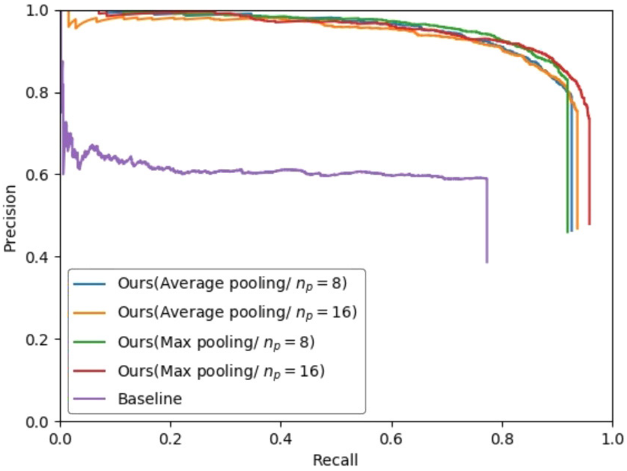


Fig. 6 Recall–precision curve for our methods and baseline method.

Table 1 Recall, precision, and average precision (AP). (Ave: average pooling, Max: max pooling, n_p: number of patches)

Method	Recall	Precision	AP
Baseline	57.3%	77.3%	46.6%
Ours (Ave./ n_p = 8)	77.5%	92.7%	88.5%
Ours (Ave./ n_p = 16)	76.3%	93.7%	90.0%
Ours (Max/ n_p = 8)	**80.7%**	91.9%	88.8%
Ours (Max/ n_p = 16)	76.5%	**95.9%**	**91.7%**

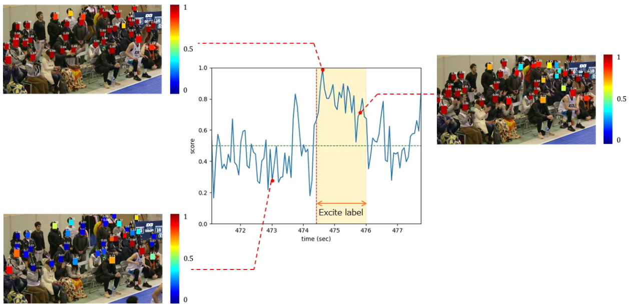


Fig. 7 Qualitative results of individual excitement. Red dotted vertical line is the moment a goal is achieved, and the following span (orange) is the moment of excitement set as a ground truth label. Score is visualized on the bounding box of each individual in each frame. Original videos are captured by Jiro Akiba/Getty Images.

(3) Individual excitement detection

Next, we conducted an individual excitement detection. In our method, the frame-level score was calculated by pooling the sigmoid scores of individual patches. Therefore, once training was completed, we obtained the individual-level score by extracting the score before pooling. Fig. 7 shows the individual- and frame-level scores in a certain period of the game. Although the scores were calculated for all individual patches, n_p was set to 8 in the training phase.

Although we could not evaluate the quantitative performance because the ground truth excitement scores of each individual were unknown, as shown in Fig. 7, our method can successfully provide the excitement score of each individual.

(4) Scene analysis

Finally, we analyzed the scene with high excitement score calculated using our method. A scene is defined as a sequence of consecutive 30 frames. To extract the scene, we first calculated the frame-level score using the excitement detection method described in Section 3. Subsequently, we calculated the scene score by the sliding window approach, which yielded the mean value for sliding the window size. The sliding window size was 30 frames, and each sliding window was non-overlapping.

We extracted scenes with the top-five highest scene scores from the second half of the final game and analyzed each scene, as shown in Table 2. We confirmed that each scene signified the goal scene from the home team by observing the video clip corresponding to the detected scene. Additionally, we discovered four scenes that were two-point scenes, which encompassed all two-point scenes in the second half of the final game. The two-point goal was the highest point goal in the 3 × 3 basketball game.

From the analysis above, we can conclude that scenes with higher excitement scores corresponded to the high-score goal scenes.

Table 2　Top-five highest scenes from calculated scene score

Rank	Frame #	Score	Point
1	24949	0.92	Two
2	22885	0.83	Two
3	21712	0.80	One
4	28469	0.80	Two
5	19741	0.79	Two

5 Conclusion

We herein proposed an object-based approach for spectator excitement detection. To evaluate the performance of our approach, we built original video datasets comprising spectators and players. Through several evaluations, we demonstrated the effectiveness of our algorithm over the traditional holistic method for our dataset and the qualitative performance of the individual excitement detection method. Using our algorithm, we analyzed the excitement scene and showed that higher excitement scores corresponded to high-point goal scenes.

References

1) D. Conigliaro, F. Setti, C. Bassetti, R. Ferrario and M. Cristani: "Attento: Attention observed for automated spectator crowd analysis", *Human Behavior Understanding*, (2013), 102-111.

2) M. Godi, R. Paolo and F. Setti: "Indirect match highlights detection with deep convolutional neural networks", *New Trends in Image Analysis and Processing – ICIAP* 2017, (2017), 87-96.

3) V. Bettadapura, P. Caroline and E. Irfan: "Leveraging contextual cues for generating basketball highlights", *Proceedings of the* 24th *ACM international conference on Multimedia*, (2016), 908-917.

4) M. Merler, D. Joshi, Q.-B. Nguyen, S. Hammer, J. Kent, J. R. Smith and R. S. Feris: "Automatic Curation of Sports Highlights using Multimodal Excitement Features", *IEEE Transactions on Multimedia*, **21** (2018).

5) R. Mehran, B. E. Moore and M. Shah: "A streakline representation of flow in crowded scenes", *European Conference on Computer Vision*, (2010), 439-452.

6) X. Wu, G. Liang, K. K. Lee and Y. Xu: "Crowd density estimation using texture analysis and learning", *2006 IEEE International Conference on Robotics and Biomimetics*, (2006), 214-219.

7) S. Wu, B. E. Moore and M. Shah: "Chaotic invariants of lagrangian particle trajectories for anomaly detection in crowded scenes", *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, IEEE*, (2010), 2054-2060.

8) J. C. S. Jacques Junior, S. R. Musse and C. R. Jung: "Crowd analysis using computer vision techniques", *IEEE Signal Processing Magazine*, **27** (2010), 66-77.

9) D. Conigliaro, P. Rota, F. Setti, C. Bassetti, N. Conci, N. Sebe and M. Cristani: "The s-hock dataset: Analyzing crowds at the stadium", *2015 IEEE Conference on Computer Vision and Pattern Recognition*, (2015), 2039-2047.

10) Simonyan Karen and Z. Andrew: "Two-stream convolutional networks for action recognition in videos", *Advances in Neural Information Processing Systems*, (2014), 568-576.

11) Robotics and Vision Laboratory, University of Minnesota, Department of Computer Science and Engineering: "A project of the Artifical Intelligence", *Monitoring Human Activity*. http://mha.cs.umn.edu/proj_events.shtml#crowd, (accessed 2020-05-29).

12) S. Ali and M. Shah: "A lagrangian particle dynamics approach for crowd flow segmentation and stability analysis", *2007 IEEE Conference on Computer Vision and Pattern Recognition*, (2007), 1-6.

13) B. Solmaz, B. E. Moore and M. Shah: "Identifying behaviors in crowd scenes using stability analysis for dynamical systems", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **34** (2012), 2064-2070.

14) S. A. Pettersen, D. Johansen, H. Johansen, V. Berg-Johansen, V. R. Gaddam, A. Mortensen, R. Langseth, C. Griwodz, H. K. Stensland and P. Halvorsen: "Soccer video and player position dataset", *Proceedings of the* 5th *ACM Multimedia*

Systems Conference, (2014), 18-23.

15) G. Waltner, T. Mauthner and H. Bischof: "Indoor activity detection and recognition for sport games analysis", *Austrian Conference on Pattern Recognition*, (2014).

16) S. Zhang, X. Zhu, H. S. Zhen Lei, X. Wang and S. Z. Li: "S3fd: Single shot scale-invariant face detector", *Proceedings of the IEEE International Conference on Computer Vision*, (2017), 192-201.

17) S. Yang, P. Luo, C. C. Loy and X. Tang: "Wider face: A face detection benchmark", *2016 IEEE Conference on Computer Vision and Pattern Recognition*, (2016), 5525-5533.

18) A. Bewley, Z. Ge, L. Ott, F. Ramos and B. Upcroft: "Simple online and realtime tracking", *2016 IEEE International Conference on Image Processing*, (2016), 3464-3468.

19) K. He, X. Zhang, S. Ren and J. Sun: "Deep residual learning for image recognition", *2016 IEEE Conference on Computer Vision and Pattern Recognition*, (2016), 770-778.

20) Jeffrey Huang: "Using two stream architecture to implement a classic action recognition method on UCF101 dataset", Github. https://github.com/jeffreyhuang1/two-stream-action-recognition, (accessed 2020-05-29).

21) A. R. Z. M. S. Khurram Soomro: "UCF101: A dataset of 101 human actions classes from videos in the wild", *arxiv preprint*, (2012).

22) C. Zach, T. Pock and B. Horst: "A duality based approach for realtime TV-L1 optical flow", *Pattern Recognition*, (2017), 214-223.

23) D. P. Kingma and J. L. Ba: "Adam: A method for stochastic optimization", *International Conference on Learning Representation*, (2015).

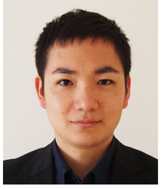
24) K. Abe, C. Nakamura, Y. Otsubo, T. Koike and N. Yokoya: "Spectator Excitement Detection in Small-scale Sports Events", *Proceedings of the 2nd International Workshop on Multimedia Content Analysis in Sports*, (2019).

阿部和広
Kazuhiro ABE
研究開発本部
数理技術研究所
Mathematical Sciences Research Laboratory
Research & Development Division

中村ちから
Chikara NAKAMURA
研究開発本部
数理技術研究所
Mathematical Sciences Research Laboratory
Research & Development Division

大坪洋介
Yosuke OTSUBO
研究開発本部
数理技術研究所
Mathematical Sciences Research Laboratory
Research & Development Division

小池哲也
Tetsuya KOIKE
研究開発本部
数理技術研究所
Mathematical Sciences Research Laboratory
Research & Development Division


横矢直人
Naoto YOKOYA
東京大学
理化学研究所
The University of Tokyo
RIKEN