Paper

Improvement of Badminton-Player Tracking Applying Image Pixel Compensation

Takashi Kamiyama†, Yoshinari Kameda†, Yuichi Ohta (member) †, and Itaru Kitahara†

Abstract Motion analysis of athletes often provides important information to improve training and strategy meetings. Visual player-tracking techniques are being developed that do not need devices. In this paper, we focus on racket sports, since they suffer from technical issues for visual tracking such as small observation size (low resolution) and large variation of player appearances. Moreover, racket sports video is usually captured by a monocular camera at a set position so that each player is observed at a top and a bottom region of the video across a net on the court. As a result, tracking accuracy is damaged by the net that often occludes players on the far side. As a solution, this paper proposes a method to improve the player-tracking accuracy in badminton video by applying an image pixel compensation technique, such as Image Inpainting. We confirm the effectiveness of our method using videos of badminton singles games.

Keywords Sports Analysis, Badminton, Monocular Camera, Player Tracking, Image Inpainting

1 Introduction

Since the motion analysis of athletes is expected to provide critical information for improving training and strategy meetings, visual player-tracking techniques are being researched and developed.[1)-4)] Since sporting matches are usually held over a long time period, the analysis time and effort can be drastically increased. On the other hand, however, assistant coaches and players want the analysis data as quickly as possible. Thus, high-speed automatic processing is critical. Putting a motion sensor on a target player makes it possible to automatically extract motion information. However, since it is impossible to ask the opposing team to wear sensors, visual tracking is expected to be a promising way to acquire positional information[5)-9)].

In this paper, we focus on badminton. For badminton analysis, we need to estimate and record such information as player position and shot type in each frame[10)]. By statistically analyzing the data, important information is provided for improving training and strategy meetings. However, the captured video contains such typical technical issues for visual tracking as small observation size (i.e., low resolution), large variation of player appearance, and partial occlusion. To easily acquire video data for on-site analysis, such data are usually captured by a monocular camera that is fixed at a relatively high place (such as a balcony) to observe all of the court's players (Fig. 1 (upper-left)). The camera needs a wide-angle lens or the distance between the camera and the target players should be long (in Fig. 1, the distance is about 25 m). As shown in Fig. 1 (bottom), the observed size of each player becomes small. Moreover, the player on the far side is partially occluded by the net on the court.

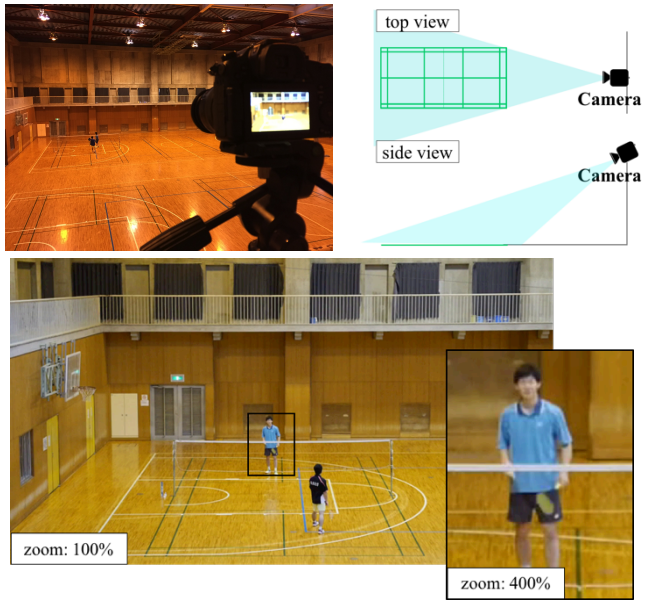


Fig. 1 Video shooting of badminton: upper-left: overview of capturing experiment; upper-right: a layout of capturing camera; bottom: example of capturing video.

Received; Revised; Accepted

†University of Tsukuba
(1-1-1 Tennodai Tsukuba, Ibaraki, Japan)

2 Related Works

Large variations of player appearances and partial occlusion are common issues not only for badminton but also for other racket sports, especially tennis, which has attracted much attention worldwide[11)-13)]. Although tennis and badminton share many characteristics, the latter's court is about three times smaller (tennis: 23.77 m×10.97 m, badminton: 13.4 m×6.1 m). As a result, occlusion (especially by the net) occurs much more often in badminton videos.

Kalal[14)] proposed a robust visual-tracking method, especially for large variations of player appearance and partial occlusion, that uses a monocular video technique called Tracking-Learning-Detection (TLD). This method executes the tracking, learning, and detection processes in parallel and compensates for each of their respective weaknesses by referring to the results of each other. In badminton video, for some badminton equipment, such as the net, the court lines are observed and the regions tend to have sharp gradients of pixel values. However, it is well known that, TLD does not work well in such region. Our proposed method solves this problem by improving the player-tracking accuracy with an image pixel compensation technique.

Image Inpainting[15)] is one of the most practical methods of image pixel compensation. Some researches[15)-17)] improve the observation quality by removing occluders. In this paper, we improve the player-tracking accuracy by removing the net region, which is one of the most troublesome occluders, using Image Painting. By applying TLD to refined video, accurate player tracking can be achieved.

3 Improving Player-Tracking Accuracy By Image Inpainting

3.1 Badminton Player Tracking by TLD

Our tracking method is based on Tracking-Learning-Detection (TLD), which accurately tracks target objects over a long time period. However, it is severely affected by the gradient of the pixel values. As a result, when the target object passes over a region where the gradient is sharp, mis-tracking is very likely. Fig. 2 shows an example of mis-tracking. In this case, two types of TLD results are shown. The left is a tracking result without a net on the court, and the right is with a net. In both images, the tracking results are denoted by colored rectangles (green: near-side player, red: far-side player). When there is no net on the court, both side players are correctly tracked; with a net, the tracking of the far-side player fails.

The reason for the mis-tracking is that part of the net region is included in the tracking-target region, because the net region is given a higher score as a tracking-target object in the learning process due to the sharp gradient. The court line region occurs a similar problem. In the regulation of badminton matches, the color of the net and the lines is strictly defined. Thus, such problems, which are caused by shape gradients, are unavoidable as long as video information is used.

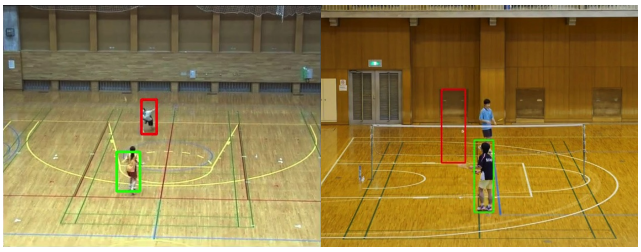
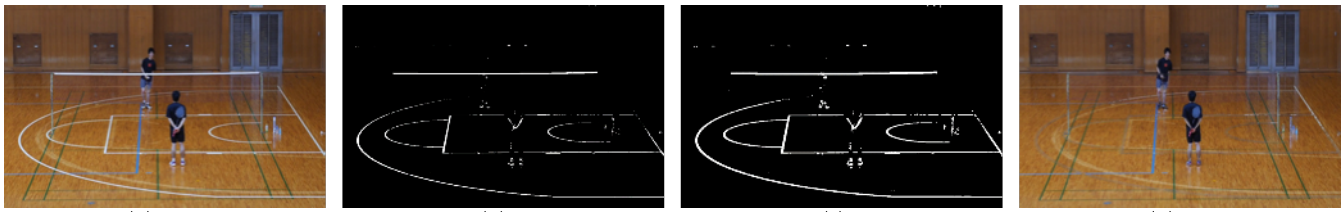


Fig. 2 TLD mis-tracking: tracking result without a net on a court (left) and with a net (right)

3.2 Image Inpainting Process for a Badminton Game Image

To solve the above problem mentioned in Section 3.1, we introduce a method combining TLD[14)] and image inpainting[15)] that reduces the shape gradient of the net and the line regions (i.e., the white region) by replacing the color of the target pixels by a color that resembles the surrounding background pixels.

Image pixel compensation processing can be categorized into two types. One fills the target pixels of the masked region by propagating the appearance of the surrounding region such as Image Inpainting [15)]. The other replaces the masked region with a region that has similar appearance with the surrounding region, such as Image Retargeting [18)]. The latter has an advantage to generate natural appearance, however if the contrast of the replaced texture was high, the gradient is still shape so that our solution does not work. On the other hand, the former does not make the gradient sharper, since it just propagates the appearance of the surrounding region where the gradient is almost flat. Since we aim to realize high speed video processing, it is another advantage of the former approach that the computational cost is smaller than the latter one.



(a) (b) (c) (d)

Fig. 3 Processing flow of Image Inpainting: (a) example of input images, (b) binarized image, (c) mask region, (d) Image Inpainting result

Figure 3 shows the processing flow of Image Inpainting. First, we extracted the white regions to which we applied Image Inpainting (Fig. 3(a)). To reduce the deleterious effects caused by the lighting condition, we converted the color space from RGB to HSV and set a threshold value in the HSV color space to binarize the white region and the other regions. Fig. 3(b) shows an example of the binarized regions. Image Inpainting does not work well when the mask region is too small. We applied the dilation process, which is one kind of morphology processing for binarized regions, extracted the mask region (Fig. 3(c)), and applied the Image Painting process to the generated mask region. Fig. 3(d) shows the Image Inpainting results. The white region that provides a sharp gradient was removed without damaging the general appearance of the input image.

3.3 Long-Term Tracking Applying Image Pixel Compensation

At the 1st frame of video sequence, TLD needs initial information about the location and size of the target players. We manually input the information by enclosing the players with bounding-boxes, which become the 1st learning templates. As mentioned in the previous section, for each frame of the input video sequence, the regions of a net of the court and lines on the floor are masked out and reduced the shape gradient by Image Inpainting. When we apply the detector of TLD to the 1st inpainted image, it finds out candidate regions of the tracking players using learned dataset. Then in the next frame, the tracker of TLD tracks the players using the inter-frame motion, which is predicted by the difference between the successive frames, to extract another learning dataset. By using the updated learned dataset, the detector finds out the players so that TLD can correct the tracking error. By repeating these processes, it is possible to realize stable long-term tracking.

4 Experiment in Gymnasiums

4.1 Video Shooting Experiments

To evaluate the effectiveness of our proposed method, we captured video sequences of badminton in a gymnasium using a monocular camera (Panasonic DMC-GH3), which can capture full HD video (1920 pixels x 1080 pixels, 30 frames/second). The shutter speed was set to 1/60 seconds. As illustrated in Fig. 1, we set the camera in a balcony to capture the badminton court by looking down in a longitudinal direction. As a result, the court net occludes the far-side players. We captured three video sequences (sequences 1, 2, 3) of practice games of men's singles by changing the color of the player's clothes. Each video was ten minutes long. We also used two video sequences of a regular game (international tournament men's singles and women's singles matches: sequences 4 and 5). We captured regular game video by a camera that apprehends 720 x 480 pixel images at 30 frames/second. Each video lasted 20 minutes, including interval times.

4.2 Evaluation of Player Detection

We confirmed the effectiveness of our proposed method by evaluating its player-tracking accuracy using the video sequences introduced in Section 4.1. TLD needs to learn the position of the target objects in the initial frame. In this experiment, we manually instructed a minimum bounding box that encloses the entire target object. The initial position is given as the center of the box's gravity.

When a bounding box (a player's candidate region) is set in an image and a player is actually observed inside it, we mark the frame as a True Positive and count them as T_P. On the other hand, when no bounding box is set in an image and there is no player in the image actually, we mark the frame as a True Negative and count them as T_N. When a bounding box (a player's candidate region) is set in an image and no player is actually observed inside it, we mark the frame as a False Positive and calculate

them as F_P. When no bounding box is set in an image and a player is actually observed in the image, we mark the frame as a False Negative and calculate them as F_N. We manually confirmed the correct marks in every frame and calculated the F-measures for the evaluation using Eqs. (1) - (3):

$$\text{Precision} = \frac{T_P}{T_P + F_P} \tag{1}$$

$$\text{Recall} = \frac{T_P}{T_P + F_N} \tag{2}$$

$$\text{F} - \text{measure} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3}$$

Figures 4(a1) and (a2) show the evaluation scores after applying TLD and TLD after Image Inpainting processing (II+TLD: our proposal) to the five above video sequences. (a1) shows the tracking results for the near-side player and (a2) for the far-side player. The mean values of Precision, Recall, and F-measure for applying only TLD were 0.54, 0.78, and 0.63, respectively. When we applied TLD after Image Inpainting processing (II+TLD: proposed), the mean values of Precision, Recall, and F-measure increased to 0.89, 0.78, and 0.83, respectively. Increasing the F-measure by 0.20 points clearly shows the effectiveness of our proposed method.

Shown as red rectangles in Fig. 4(b1), TLD mis-tracked the net region as a far-side player due to occlusion. On the other hand, our proposed method (II+TLD) accurately tracked the far-side player by solving the occlusion problem (Fig. 4(b2)).

However, when we tracked the near-side player in sequence 3, the F-measure value drastically decreased with II+TLD. In this sequence, since the near-side player is in white, the Image Inpainting process removed the region with the net and line regions. Since clothing colors can be known in advance, we believe it is possible to avoid such mis-decision using other information such as the shape of the white regions.

4.3 Evaluation for Player Tracking

In Section 4.2, we described how our proposed method detected a player better than TLD throughout all the frames of all of the sequences. However, as shown in Figs. (c1) and (c2), there are some frames when our proposed method (II+TLD) degraded the tracking result more than TLD. Here, we evaluate our method based on improvement and worsening rates. Each value can be calculated by Eqs. (4) and (5). The total number of frames is K. The arrow ($\rightarrow$) in the equation denotes that our method changes the states given by TLD to different states. Each state (T_P, T_N, F_P, F_N) is the same as the one introduced in the previous section:

$$\text{Improvement rate} = \frac{(\{F_P, F_N\} \rightarrow \{T_P, T_N\}) + (\{F_P\} \rightarrow \{F_N\})}{2K} \tag{4}$$

$$\text{Worsening rate} = \frac{(\{T_P, T_N, F_N\} \rightarrow \{F_P\}) + (\{T_P, T_N\} \rightarrow \{F_N\})}{2K} \tag{5}$$

Basically, the improvement rate depends on the number of frames in which a false decision is revised to a true one by applying our method, and the worsening rate is exclusive of the improvement rate. However, when the decision is False Negative, the mis-tracking result can be complemented by referring to the decision of the back and forth frames. So we counted the frames $\{F_P\}\rightarrow\{F_N\}$ in which the False-Positive decision is revised to a False Negative by applying our method to the improvement rate.

Table 1 shows the results of the improvement and worsening rates for each video sequence. The mean value of the improvement rate for all the sequences is 0.39, and the worsening rate is 0.08. Our proposed method works well in almost all of the frames. One reason for the worsening is that the court line is not white. In this case, the line region has a sharp gradient so that TLD incorrectly learned the foreground (target object), including such background regions as lines. We must avoid this mis-learning by flexibly defining the color of the lines depending on the captured environment.

Table 1 Improvement and worsening rate for each sequence

	Improvement rate	Worsening rate
sequence 1	0.21	0.10
sequence 2	0.06	0.06
sequence 3	0.28	0.24
sequence 4	0.74	0.04
sequence 5	0.43	0.05

5 Conclusion

We proposed a method to improve the accuracy of visual player tracking using such image compensation processing as Image Inpainting. Our research target was racket sports since they suffer from such technical issues as small observation size (low resolution) and large variation of player's appearance. Racket sports video is usually captured by a monocular camera at a regular position to observe each player at the top and bottom regions of the video across a net on the court. As a result, tracking accuracy is damaged by the net that often occludes players on the far side. We confirmed the effectiveness of our method using videos of badminton single games.

This work was partially supported by JSPS KAKENHI Grant Number 25280056.

References

1) S. Kawamura, T. Fukusato, T. Hirai, and S. Morishima: "Efficient Video Viewing System for Racquet Sports with Automatic Summarization Focusing on Rally Scenes," 41th Int. Conf. & Exhibition on Comput. Graphics & Interactive Techniques (SIGGRAPH) Posters, 62, (July 2014)
2) T. Lan, L. Sigal, and G. Mori: "Social Roles in Hierarchical Models for Human Activity Recognition," IEEE Conf. Comput. Vision & Pattern Recognition (CVPR), pp. 1354-1361 (June 2012)
3) W-L. Lu, J-A. Ting, K. P. Murphy, and J. J. Little: "Identifying Players in Broadcast Sports Videos Using Conditional Random Fields," IEEE Conf. Comput. Vision & Pattern Recognition (ICCV), pp. 3249-3256 (June 2011)
4) F. Yoshikawa, T. Kobayashi, K. Watanabe, and N. Otsu: "Automated Service Scene Detection for Badminton Game Analysis Using CHLAC and MRA," Int. J. Comput., Electrical, Automation, Control & Inf. Engineering, **4**, 2, pp. 331-334 (Feb. 2010)
5) S. Duffner and C. Garcia: "PixelTrack: A Fast Adaptive Algorithm for Tracking Non-rigid Objects," IEEE Int. Conf. Comput. Vision (ICCV), pp. 2480-2487 (Dec. 2013)
6) H. B. Shitrit, J. Berclaz, F. Fleuret, and P. Fua: "Tracking Multiple People under Global Appearance Constraints," IEEE Conf. Comput. Vision & Pattern Recognition (ICCV), pp.137-144 (Nov. 2011)
7) K. Fragkiadaki and J. Shi: "Detection Free Tracking: Exploiting Motion and Topology for Segmenting and Tracking under Entanglement," IEEE Conf. Comput. Vision & Pattern Recognition (CVPR), pp. 2073-2080 (June 2011)
8) J. Liu, P. Carr, R. T. Collins, and Y. Liu: "Tracking Sports Players with Context-conditioned Motion Models," IEEE Conf. Comput. Vision & Pattern Recognition (CVPR), pp. 1831-1837 (June 2013)
9) S. Chen, A. Fern, and S. Todorovic: "Multi-object Tracking via Constrained Sequential Labeling," IEEE Conf. Comput. Vision & Pattern Recognition (CVPR), pp. 1130-1137 (June 2014)
10) S. L. Teng and R. Paramesran: "Detection of Service Activity in a Badminton Game", IEEE Region 10 Conf. (TENCON), pp. 312-315, (Nov. 2011)
11) F. Yan, J. Kittler, D. Windridge, W. Christmas, K. Mikolajczyk, S. Cox, and Q. Huang: "Automatic Annotation of Tennis Games: An Integration of Audio, Vision, and Learning", Image & Vision Computing, **32**, 11, pp. 896-903 (Nov. 2014)
12) M-Y. Fang, C-K. Chang, N-C. Yang, C-M. Kuo, and S-K. Guang: "Robust Player Tracking for Broadcast Tennis Videos with Adaptive Kalman Filtering," J. Inf. Hiding & Multimedia Signal Process., **5**, 2, pp. 242-262 (Apr. 2014)
13) N. Maruyama and K. Fukui: "Motion Analysis for Broadcast Tennis Video Considering Mutual Interaction of Players," 12th Conf. Machine Vision Applications (MVA), pp. 454-458 (June 2011)
14) Z. Kalal, K. Mikolajczyk, and J. Matas: "Tracking-Learning-Detection," IEEE Trans. Pattern Analysis & Machine Intelligence (PAMI), **34**, 7, pp. 1409-1422 (July 2012)
15) M. Bertalmio, G. Sapiro, V. Caselles, and C. Ballester: "Image Inpainting," 27th Int. Conf. & Exhibition on Comput. Graphics & Interactive Techniques (SIGGRAPH), pp. 417-424 (July 2000)
16) A. Telea: "A Image Inpainting Technique Based on the Fast Marching Method," J. Graphics Tools, **9**, 1, pp. 23-34 (May 2004)
17) K. A. Patwardhan, G. Sapiro, and M. Bertalmio: "Video Inpainting Under Constrained Camera Motion," IEEE Trans. Image Process., **16**, 2, pp. 545-553 (Feb. 2007)
18) D. Vaqueroa, M. Turka, K. Pullib, M. Ticob, N. Gelfandb, "A Survey of Image Retargeting Techniques", Proc. SPIE 7798, Applications of Digital Image Processing XXXIII, 779814 (Sep. 2010);

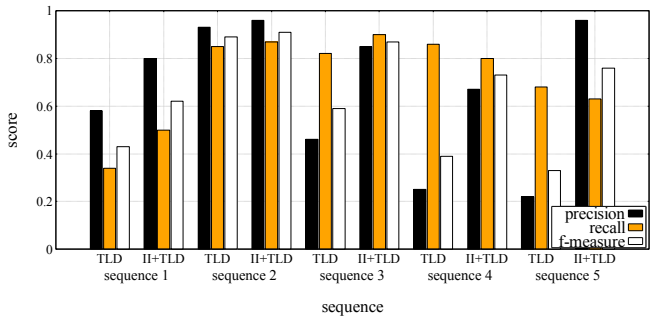
Takashi Kamiyama
received his B.E. and M.E. degrees in System Engineering from University of Tsukuba, Japan in 2014 and 2016, respectively. His research interests are computer vision.

Yoshinari Kameda
received his B.E and M.E and Ph.D rom Kyoto University in 1991, 1993, and 1999. He had a faculty position at Kyoto University in 1999-2003. He was a visiting scholar at MIT in 2001-2002. In 2003 he joined the University of Tsukuba and he is a professor at University of Tsukuba. His research interests include the enhancement of human vision, augmented reality, mixed reality, video media processing, computer vision, and sensor fusion.

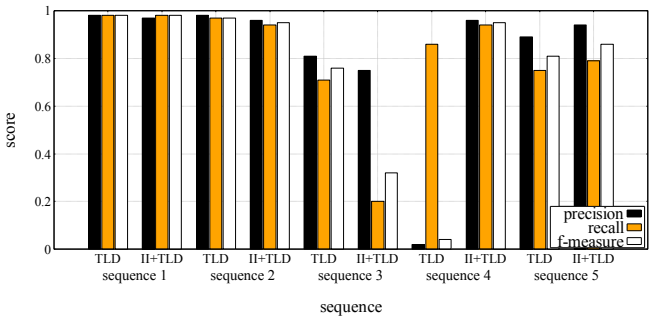
Yuichi Ohta
received his B.E. and M.E. degrees in Engineering from Kyoto University, Japan in 1972 and 1974, respectively. He received Ph.D. from Kyoto University in 1980. 1978-1981, he was Research Associate of Kyoto University. 1981-1987 Assistant Professor at the University of Tsukuba. 1987-1992 Associate Professor and 1992-2004 Professor at the University of Tsukuba. 2004-2009

Professor in the Graduate School of the University of Tsukuba. 2009-2012 he served as Provost of the Graduate School. 2013-2014, he was Vice President of the University of Tsukuba. 1982-1983, he was Visiting Scientist in Computer Science Department, Carnegie Mellon University. He is an IAPR Fellow, an IEICE Fellow, and an IPSJ Fellow.

Itaru Kitahara
received his B.E. and M.E. degrees in Science Engineering from University of Tsukuba, Japan in 1994 and 1996, respectively. In 1996, he joined Sharp Corporation. 2000-2003, he was a research associate of University of Tsukuba. He received his PhD in 2003. 2003-2005, he was a researcher at ATR. 2005-2008, he was an assistant Professor at the University of Tsukuba. Since 2008, he has been an assistant professor at the University of Tsukuba. His research interests include computer vision, mixed reality, and intelligent image media.



(a1)



(a2)

(b1)

(b2)

(c1)

(c2)

Fig. 4 Experimental results of visual player tracking: (a1) tracking results for near-side player, (a2) for far-side player, (b1) TLD mis-tracks net region as a far-side player due to occlusion problem, (b2) our proposed method (II+TLD) accurately tracks far-side player by solving occlusion problem, (c1) (c2) our proposed method (II+TLD) makes the tracking result worse than TLD.