EasyChair Preprint

№ 2677

Action Recognition in Sports Video Considering Location Information

Rina Ichige and Yoshimitsu Aoki

February 15, 2020

Action Recognition in Sports Video Considering Location Information

Rina Ichige[1] , Yoshimitsu Aoki[1]

[1] Keio University, 3-14-1, Hiyoshi, Kohoku-ku, Yokohama 223-8522, Japan
richige@aoki-medialab.jp

Abstract. The purpose of this study is to develop a tactics analysis system using image recognition for rugby. With the Rugby World Cup in 2019 and the Tokyo Olympics in 2020, demand for sports video analysis is increasing. Rugby has more complicated play such as dense play than other sports, and the ball is hidden between players, making it difficult to track. By developing a high-precision analysis technology for rugby with few research cases, we thought that it could be used for other sports and industrial fields other than sports. In this research, we propose a method that adds spatial information to time-series information as a new feature. Using the coordinates obtained by projectively transforming the match video onto the bird's-eye view image, play classification was performed using the player position, the ball position, and the dense area position as feature amounts. Also, in order to further improve the detection accuracy of the boundaries between plays, attention was paid to the positional relationship of each player on the field.

Keywords: Dense play, heatmap features, Subdivision of play area.

1 Introduction

In recent years, the demand for video analysis utilizing ICT (Information and Communication Technology) as content for strengthening teams and tactics and watching sports has been increasing in the sports world. In particular, companies that have had little relevance in the sports field have begun to actively participate in the event, especially since the 2019 Rugby World Cup Japan Games in Japan and the 2020 Tokyo Olympics have hosted global sports festivals.

For team sports such as soccer and basketball, action recognition for the actions of players during a game has already been performed. On the other hand, there are few studies on rugby at present. The reason is the rugby's playing characteristics. First, the number of players participating in the game is 15 per team, which is larger than volleyball (6), baseball (9), soccer (11), and football (11). In addition, since the movements and postures of the players vary widely, it is necessary to pay attention to the movements of each player, such as the speed and direction in which they run. In addition, players such as “Scrum”, “Lineout” etc. are often hidden behind shadows due to contact play or dense play, which makes tracking difficult. Play continues until a goal is scored,

the ball goes out of the touch line, or an offense occurs, during which time offense and defense changes occur frequently.

Therefore, by developing a highly accurate analysis technology for rugby with complex game characteristics, it was thought that it would be possible to use it for other sports and use it as content for watching games on TV.

In addition, due to the lack of widespread analysis technology, rugby play is currently tagged manually by analysts after watching the match video (40 minutes in the first half). We thought that if we could automatically extract the necessary play scenes, we would reduce the burden on analysts and directly use the time for tagging for on-site coaching of players, which would strengthen the team.

2 Previous Work

Conventionally, a play classification method utilizes an LSTM (Long Short-Term Memory)[1] for handling sequence data. As shown in Fig. 1, a still image cut out from a video is converted into a feature amount by a CNN (Convolutional Neural Network) [2] for each frame, and is input to an LSTM with 512 nodes in a fully connected layer. This is to calculate the probability distribution of the labeled play by averaging the outputs of the nodes of the number of classes in all the connected layers [3]. This method has the advantage that everything from video design to feature classification to play classification can be performed automatically by machine learning, but since all the play is detected by inputting the frame images all at once. The accuracy of each play is not high.

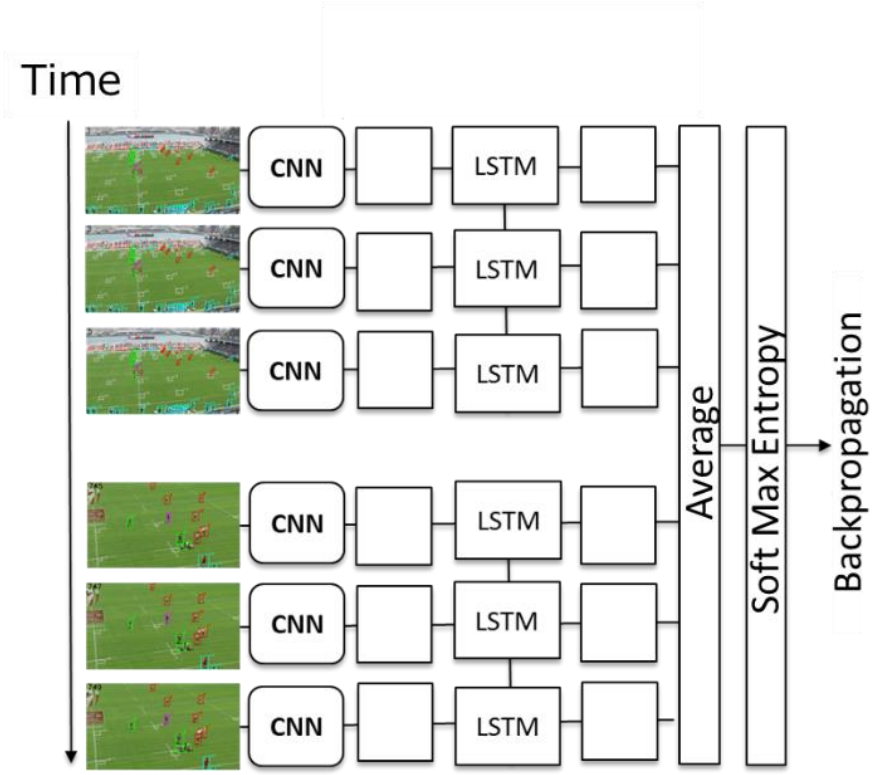

Fig. 1. Previous work(play estimation using LSTM)

3 Proposed Method

This time, we focused on the characteristics of rugby, where players and balls are easily hidden in dense areas, clarified the relationship between players and players and between players and balls, and focused on the role of each player by subdividing the play area. At this point, we conducted research with a view to watching TV broadcast content by automatic detection and real-time detection of play scenes required for analysis. In this study, since the player's play recognition is the main axis, we analyzed the game video in which the player position, ball position, and dense area position were taught in advance. For the feature extraction, three methods were considered: the handcraft feature, the heatmap feature, and the use of the subdivision region.

3.1 Shooting method

As a method of detecting a player's position, it is conceivable to take a picture by switching the multi-view camera, but this time we will use the video taken by the fixed camera in terms of installation cost.

In the conventional method, since an image obtained by pan shooting with only one camera is input, there is a problem that fluctuations in camera work when the image is taken affect the movement of players. there were.

Therefore, this time, we fixed the camera and considered a shooting method that shows the whole field. In this method, if a player can be detected, the position of the player with respect to the field is fixed, so that the positional relationship can be accurately set as a feature amount.

In order to detect even distant players more accurately, images taken using two cameras for the left and right fields were merged, as shown in Fig. 2.

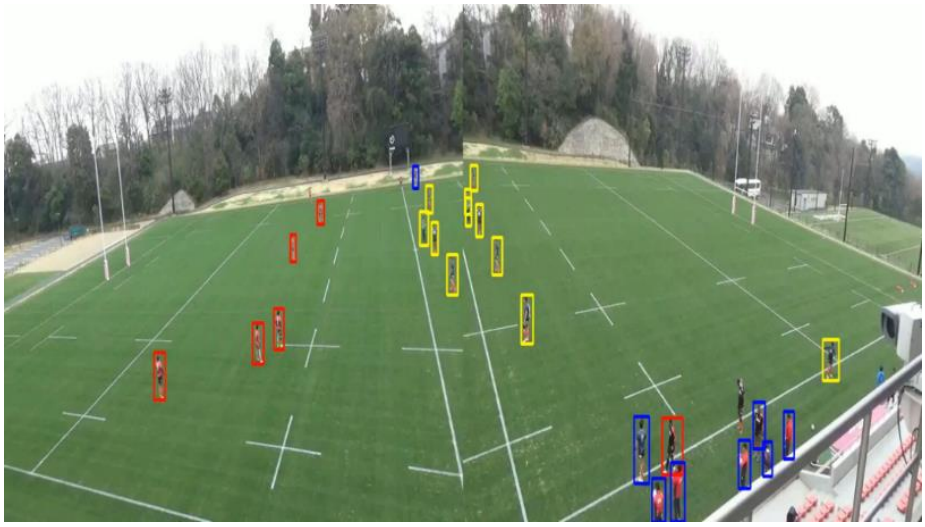

Fig. 2. Fixed shooting (left and right merge)

3.2 Conversion to overhead image

In order to design features that are not affected by camera work, the projection transformation matrix (1)[4][5][6] was estimated for each frame from the movement of the corresponding points (4 points) (x, y) on the white line, and converted to an overhead image Player / ball positions (x ', y ') (Fig. 3) were used. As a result, the distance between the players does not depend on the distance between the object and the camera, and the positional relationship between the players is uniformly maintained.

$$\begin{pmatrix} x' \\ y' \\ 1 \end{pmatrix} = \begin{pmatrix} a_1 & b_1 & c_1 \\ a_2 & b_2 & c_2 \\ a_0 & b_0 & c_0 \end{pmatrix} \begin{pmatrix} x \\ y \\ 1 \end{pmatrix} \quad (1)$$

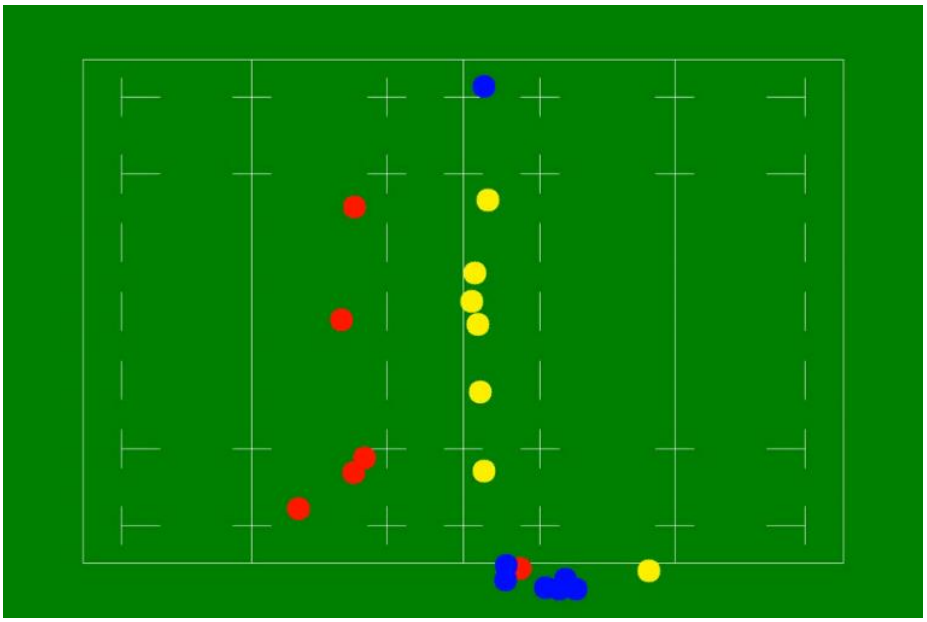

Fig. 3. Conversion to overhead image

3.3 Handcraft features

In this method, handcraft features (static features and dynamic features) were designed and input for each frame instead of an image as shown in Fig. 4. The feature that did not automatically extract features using CNN as in the conventional method of Fig. 1 is that there are few precedents in rugby research, and at this time a large number of videos with the player and ball positions taught are prepared. Was difficult. If high-dimensional data such as an image is used as an input for a small amount of learning data, learning may be affected by unnecessary information included in the image, and over-learning may occur.

The average of the player positions and the variance of the player positions were used for the static features. The average player position was determined for all players in that frame without distinguishing between teams. Then, the variance of the player positions was calculated from the average position obtained by the following equation.

$$s_x{}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \quad (2)$$

$$s_y{}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 \quad (3)$$

The average of the player velocities between certain frames was used as the dynamic feature value. This time, since it is used as a feature value for classification of play regardless of the team, the sum of the speeds issued for each player with the same ID is fully added without distinguishing the team, and it is detected in that frame that both teams combined By normalizing with the number of all players performed, the feature amount in the sense of the average of the movement amount and direction of all the players between l frames was obtained. For example, in dense play such as “Lineout”, the speed of the player is small in the dense area, so the magnitude of the speed is small, and in “Turnover”, the change of offense and defense occurs between the teams, so it is the moment when the overall speed vector is reversed. Conceivable.

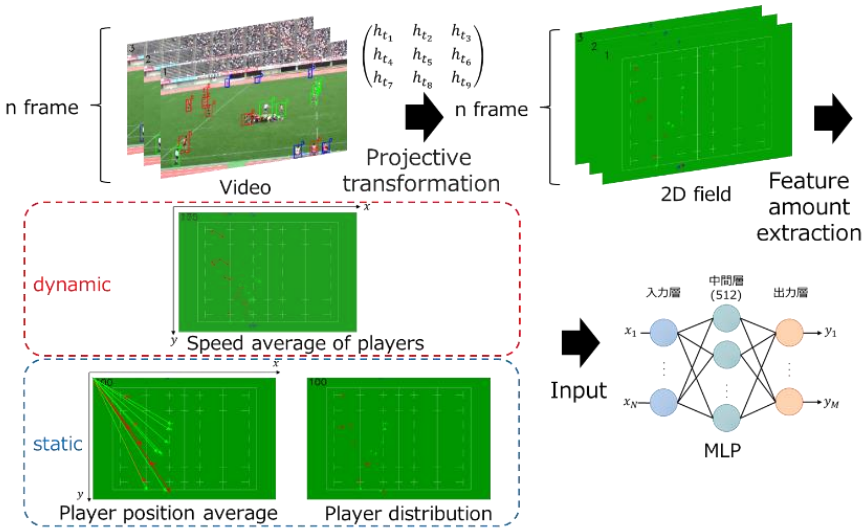

Fig. 4. Handcraft features

3.4 Heatmap features

The handcraft feature is a high-dimensional image called a low-dimensional matrix containing numerical values for each time series. On the other hand, the heatmap feature quantity has a feature that only necessary information is extracted while maintaining the format of an image as shown in Fig. 5. Since the size of the rugby field varies depending on the shooting location, the input frame is first normalized. Like the handcraft feature, the feature includes a static feature and a dynamic feature.

A zero matrix (black image) having the same size as the normalized bird's-eye view image (Y, X) is prepared for 6 channels (4 static feature values + 2 dynamic feature values) × the number of frames.

For the coordinates at which the position is detected, a pixel value of 1 is inserted instead of 0 in the static feature amount. A Gaussian filter is applied to this image. This

operation is performed for all four channels $ch_0 \sim ch_3$ (the player position p_{ayx} of the team 1, the player position p_{byx} of the team 2, the ball position b_{yx}, and the dense area position c_{yx}).

In the case of the dynamic feature amount, the velocity vector (ch_4, ch_5) corresponding to the movement compared with the past frame (y_{prev}, x_{prev}) is inserted.

$$ch_0 = \begin{pmatrix} p_{a00} & \cdots & p_{a0X} \\ \vdots & \ddots & \vdots \\ p_{aY0} & \cdots & p_{aYX} \end{pmatrix} \tag{4}$$

$$ch_1 = \begin{pmatrix} p_{b00} & \cdots & p_{b0X} \\ \vdots & \ddots & \vdots \\ p_{bY0} & \cdots & p_{bYX} \end{pmatrix} \tag{5}$$

$$ch_2 = \begin{pmatrix} b_{00} & \cdots & b_{0X} \\ \vdots & \ddots & \vdots \\ b_{Y0} & \cdots & b_{YX} \end{pmatrix} \tag{6}$$

$$ch_3 = \begin{pmatrix} c_{00} & \cdots & c_{0X} \\ \vdots & \ddots & \vdots \\ c_{Y0} & \cdots & c_{YX} \end{pmatrix} \tag{7}$$

$$\begin{pmatrix} ch_4 \\ ch_5 \end{pmatrix} = \begin{pmatrix} \begin{pmatrix} y_{00} & \cdots & y_{0X} \\ \vdots & \ddots & \vdots \\ y_{Y0} & \cdots & y_{YX} \end{pmatrix} \\ \begin{pmatrix} x_{00} & \cdots & x_{0X} \\ \vdots & \ddots & \vdots \\ x_{Y0} & \cdots & x_{YX} \end{pmatrix} \end{pmatrix} - \begin{pmatrix} \begin{pmatrix} y_{prev00} & \cdots & y_{prev0X} \\ \vdots & \ddots & \vdots \\ y_{prevY0} & \cdots & y_{prevYX} \end{pmatrix} \\ \begin{pmatrix} x_{prev00} & \cdots & x_{prev0X} \\ \vdots & \ddots & \vdots \\ x_{prevY0} & \cdots & x_{prevYX} \end{pmatrix} \end{pmatrix} \tag{8}$$

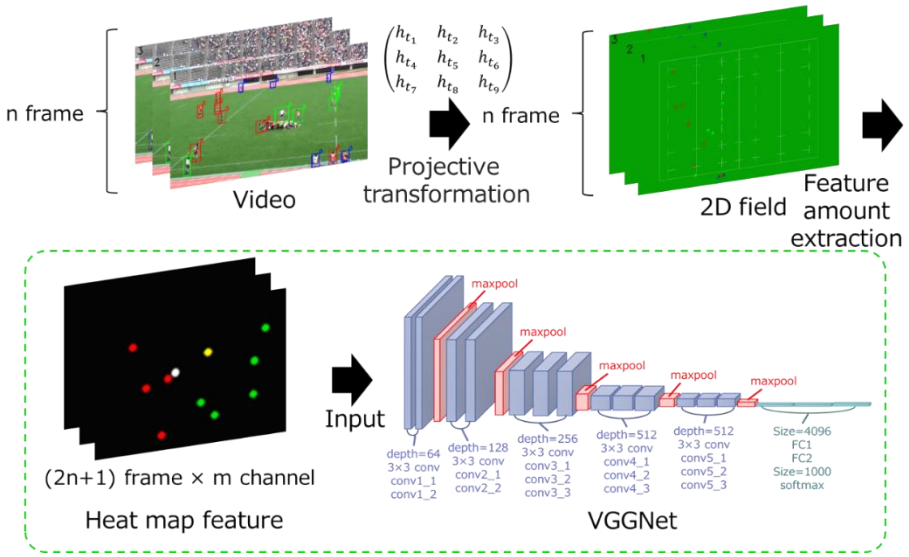

Fig. 5. Heatmap features

3.5 Use of heatmap features and subdivision areas

In EPV (expected possession value) [7], the expected value of the score is calculated by the dispersion of players on the basketball court. In RPR (Reachable polygonal region) [8], the dominant region of each player is calculated using approximate bisectors drawn for each pair of players. Referring to these conventional studies, consider where the players, balls, and dense areas belong to the subdivided areas as shown in Fig. 6. In the heatmap feature, the variance of players, the ball position, and the position of the dense area in the entire field were applied on the heatmap. For example, in the case of "Lineout", a dense area is assigned to an area near the line, so that it is possible to distinguish it from "Scrum", which is the same dense play. In Fig. 6, the location of the average player position / ball position / dense area position is labeled in the subdivided play area, and added as a new channel to heatmap feature amounts (4) to (8). .

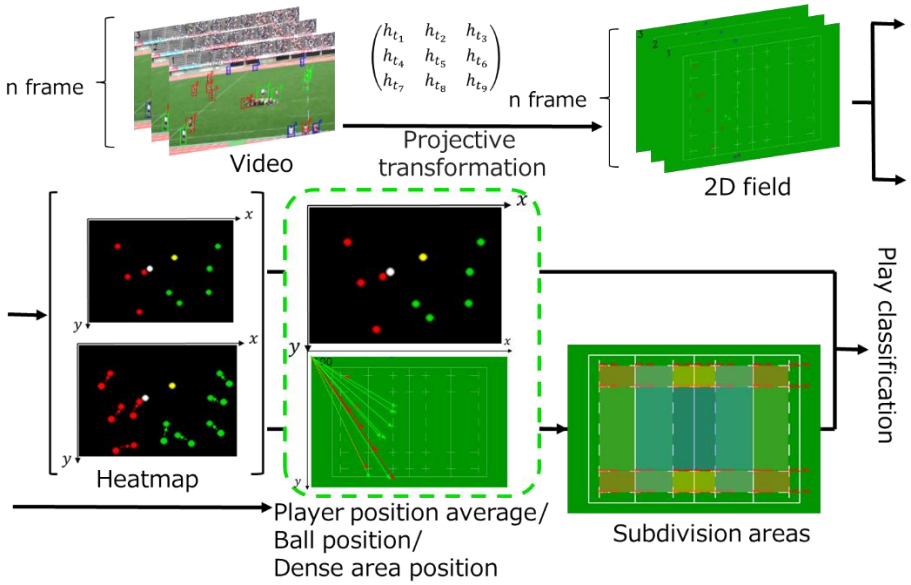

Fig. 6. Use of heatmap features and subdivision areas

4 Experiment

4.1 Dataset details

As shown in Table 1, the data set used in this experiment is a set of eight fixed images (30 fps, Fig. 2) obtained by merging two fixed cameras into a left-right image. Used as a minute video.

Play labels were given for six classes as shown in Table 2. In fact, 7 classes are labeled, including the other play “Other Play”. However, as shown in Table 2, the proportion of “Other Play” included in the data set is large, For the purpose of improving the accuracy of play classification in this study, this time we used labels for 6 classes excluding "Other Play" in order to consider the difference in accuracy due to the method.

Table 1. Dataset

movie	1	2	3	4	5	6	7	8
frame (7label)	21177	19779	1919	21536	22239	44968	37121	40129
frame (6label)	1790	1589	1350	770	764	2073	1313	2048

Table 2. label

label	0	1	2	3	4	5
play	Scrum	Lineout	Kick off	Kick counter	Turnover	Penalty

4.2 Evaluation index

In Table 3, Accuracy, Precision, Recall, and F-measure were used as evaluation indices in Table 3.

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \tag{9}$$

$$Precision = \frac{TP}{TP+FP} \tag{10}$$

$$Recall = \frac{TP}{TP+FN} \tag{11}$$

$$F\text{-}measure = \frac{2 \cdot Recall \cdot Precision}{Recall+Precision} \tag{12}$$

Table 3. evaluation index

		GT	
		P	N
Predict	P	TP	FP
	N	FN	TN

5 Discussion

"Handcraft" method using handcraft feature amount, "heatmap" method using only heatmap feature amount, and "heatmap" indicating where the average player position, ball position, and dense area position belong to the subdivided play area in Fig. 6 Compare each method as "player", "ball", "crowd", and unify all of them as all. GT is above, prediction results are below, light blue is "Scrum", orange is "Lineout", gray is "Kick off", yellow is "Kick counter", blue is "Turnover", green is "Penalty".

In this article, we consider video 6, the labels included "Scrum" 7.5%, "Lineout" 14%, "Kick off" 31%, "Kick counter" 18%, "Turnover" 18%, and "Penalty" 12%.GT and predicted values are shown in Figs. 7 to 12.

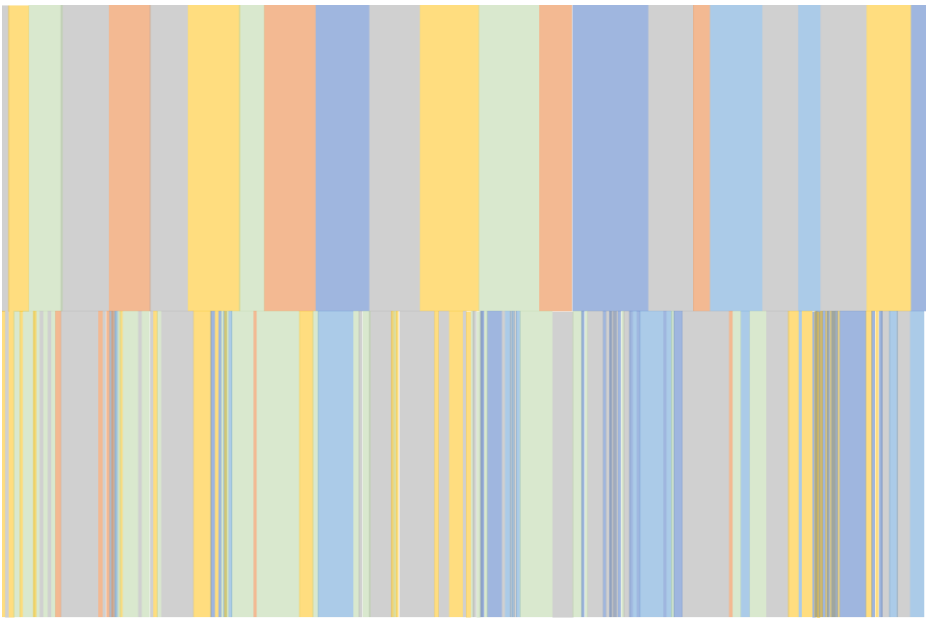

Fig. 7. GT and predict(handcraft)

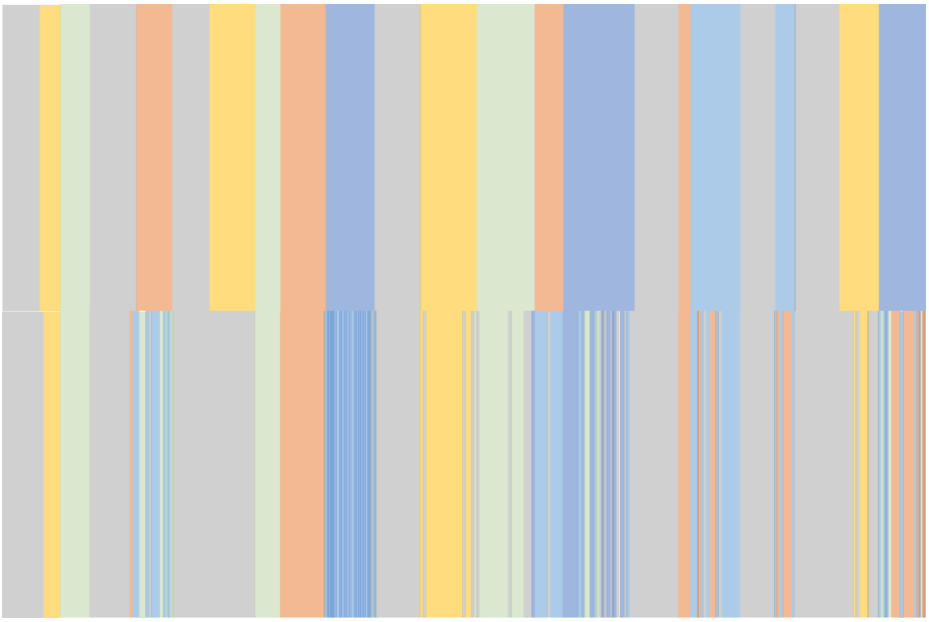

Fig. 8. GT and predict(heatmap)

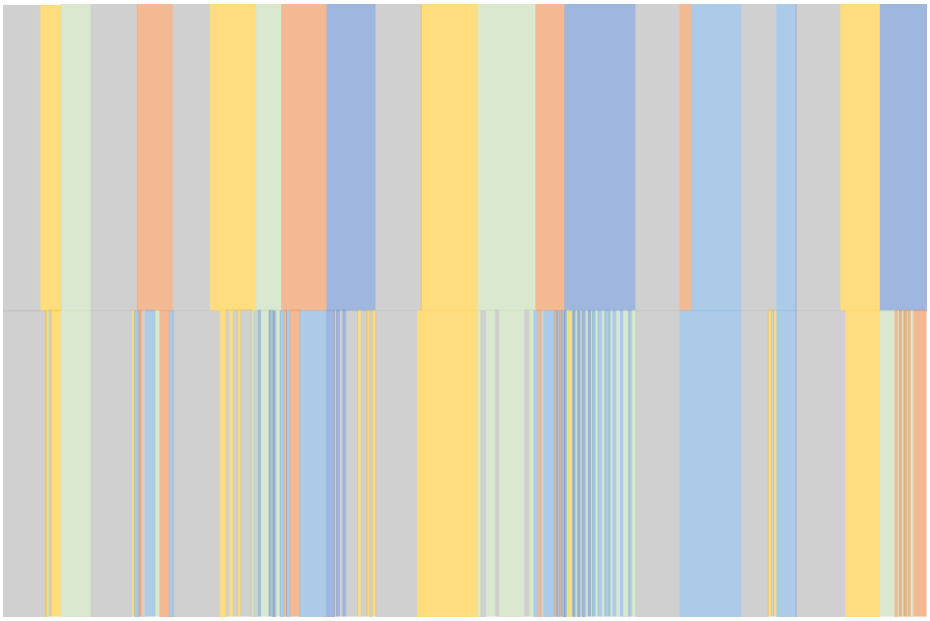

Fig. 9. GT and predict(player)

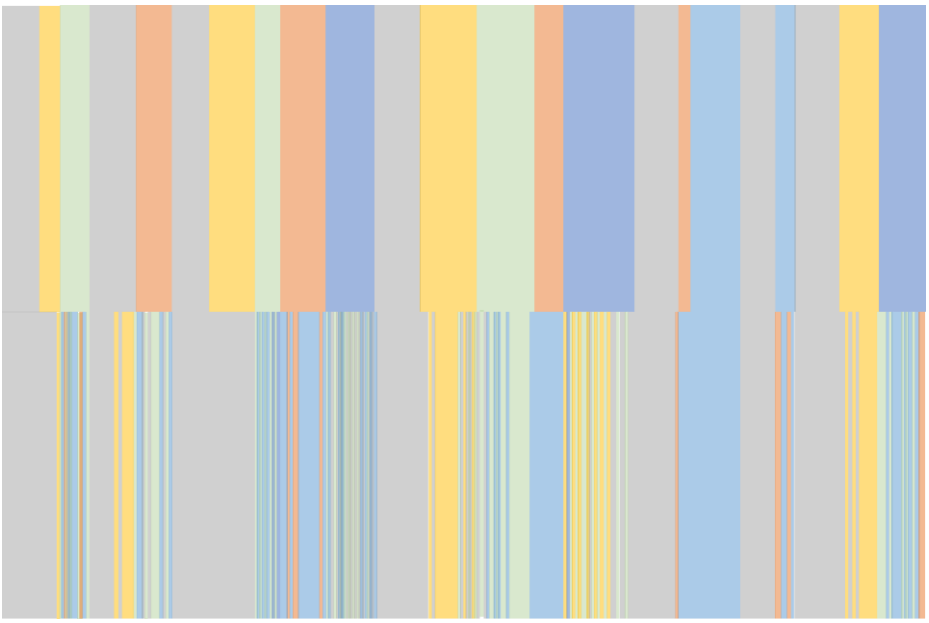

Fig. 10. GT and predict(ball)

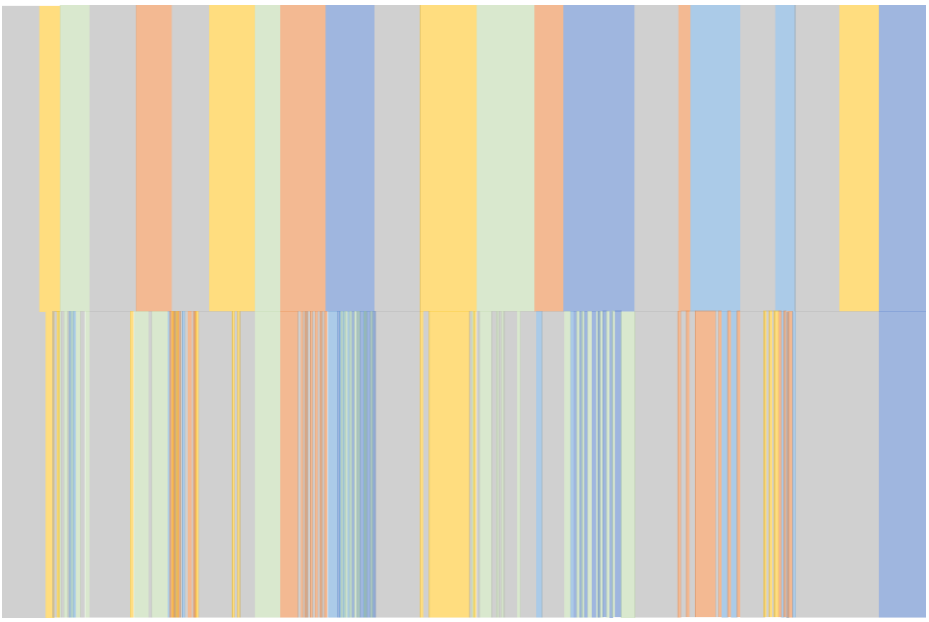

Fig. 11. GT and predict(crowd)

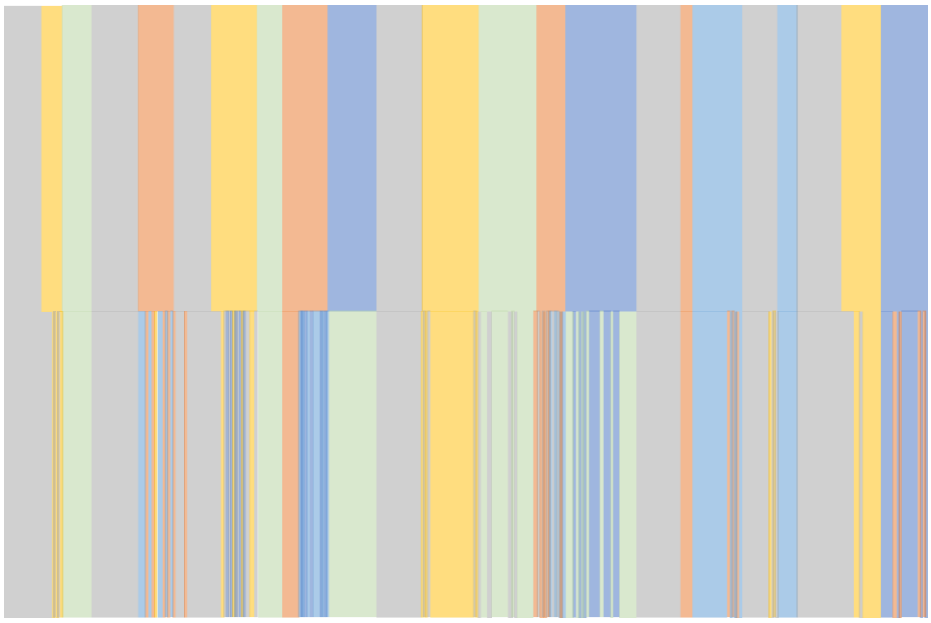

Fig. 12. GT and predict(all)

5.1 Handcraft features and Heatmap features

From Fig. 7, Fig. 8, when “heatmap” is used in “Scrum”, the “Kick off”, “Kick counter”, “Turnover”, “Penalty” and false detection (FN) in “handcraft” have been improved, but “Lineout” has been falsely detected. (FN). This is thought to be due to the fact that the frequency of switching between plays in GT is high, and the part that was difficult to appear with the scalar amount of handcraft is easy to distinguish “Scrum” as a play with small dispersion in “heatmap”.

When using “heatmap” in “Lineout”, the part that was falsely detected (FN) as “Kick off”, “Kick counter”, and “Penalty” in “handcraft” has been improved, but it is easier to falsely detect (FN) as “Scrum”. This is thought to be because, as with “Scrum”, the “heatmap” makes it easy to distinguish small play from play that is not, but dense play with small variance is unlikely to appear as a difference in feature amount.

Regarding “Kick off”, when using “heatmap”, the part (FN) that was incorrectly detected as “Scrum” and “Turnover” in “handcraft” has been improved, and the recall has been increased, but the “Kick counter” has been mistakenly detected as “Kick off” (FP) and it has become easier to falsely detect (FP) I have.

Also, using “heatmap" for the “Kick counter”, the part that was falsely detected as “Penalty” in “handcraft” (FN), the part where “Scrum”, “Lineout”, “Kick off” was falsely detected as “Kick counter” (FP) has been improved, and Recall, Precision is higher. This is probably because the variance clearly appears as a heatmap feature and is easily distinguished from a play (“Scrum”, “Lineout”) with a small variance, but the “Kick off” cannot be distinguished from the “Kick counter” performed at the end of the field.

When using “heatmap” in “Turnover”, the part that was falsely detected as “Scrum” in “handcraft” (FN), the part where “Kick off”, “Penalty” was falsely detected as “Turnover” (FP) has been improved, and the Recall and Precision are higher, It is easy to false detection (FN) with “Lineout”. This is because the change in heatmap feature amount is more likely to appear as a variance than “handcraft”, but “Turnover” that occurs before and after the “Kick counter” is less likely to appear as a difference in heatmap feature amount. It is considered that the variance in is influenced by “Lineout”.

5.2 Subdivision of play area

In “Scrum”, from Fig. 8 and Fig. 9 in Recall, in the “heatmap”, the part that was falsely detected as “Lineout” (FN) has been reduced in the “player”. This is because in “Scrum” and “Lineout” where GT transitions from “Lineout” to “Scrum”, both “Scrum” and “Lineout” are dense play, so the variance in “heatmap” is similar, but by using “player”, players are concentrated in the “Lineout” where players concentrate on the line and in areas other than the line It is thought that the distinction of concentrated “Scrum” is clarified.

In “Lineout”, from Fig. 8, Fig. 9, Fig. 12 in Recall, in “heatmap”, the part that was falsely detected as “Penalty” (FN) was reduced in “all”, but it is easier for “player” to falsely detect “Scrum” and “Kick off”. This is considered to be because most of “Penalty” in GT occurs before and after “Lineout”, but by using “crowd”, it becomes clear to distinguish between “Lineout” where dense areas are detected at the line and “Penalty” where dense areas are not detected. Also, in the “player”, the “Lineout” that occurs before and after “Scrum” in GT is similar in the area where players concentrate, and the “Lineout” that occurs before and after “Scrum” is both players concentrated on the line, making it difficult to distinguish .

In “Kick off”, from Fig. 8, Fig. 9, Fig. 10 in Recall, in the “heatmap”, the part that was falsely detected (FN) as “Kick counter” and “Turnover” is reduced in the “player”, but it is easier to falsely detect as “Lineout” (FN) in “crowd”. This happens in GT before and after “Kick off” for both “Kick counter” and “Turnover”, but “Kick off” where the area where players concentrate by “player” is on the line and “Kick counter” where the area where players concentrate is other than line It is thought that the distinction of “Turnover”, whose formation changes instantaneously, becomes clearer. Also, in GT, most of “Lineout” occurs before and after “Kick off”, but it is affected by “Lineout” at the boundary between “Kick off” and “Lineout”, and it is difficult to distinguish because dense areas are detected at the line side Conceivable.

In the “Kick counter”, from Fig. 8, Fig. 10, Fig. 12 in Recall, the part that was falsely detected as “Kick off” and “Penalty” (FN) in “heatmap” decreased in “all”, but increased in “ball”. This is a “Kick counter”, “Penalty” occurs before and after “Kick off” in GT, but “Kick off”, where the area where players concentrate by the “player” is on the line, and “Kick counter”, where the area where players concentrate is other than the

line, This is probably because the distinction between “Penalty”, where the concentration area is not limited, is clarified. In addition, it is considered that the frequency of transition in a short frame interval from kick to kick to “Penalty” in GT is high, and the ball position is affected by the play before and after, making it difficult to distinguish.

In “Turnover”, from Fig. 8, Fig. 10, Fig. 11 in Recall, in “heatmap”, the false detection of “Scrum”, “Lineout”, “Kick off”, “Penalty” (FN) is reduced in the “crowd”, but in the “ball”, “Scrum”, “Kick off”, “Kick counter”, “Penalty” is more likely to be falsely detected. This is due to the fact that most of “Scrum” and “Turnover” in GT immediately before “Lineout” most of “Penalty”, but crowded area is detected in places other than line by “crowd”, “Scrum”, dense area is detected in line It is thought that the distinction between “Lineout” and “Turnover” where the dense area is not detected is clarified, and it is clear that the distinction between “Penalty” and “Turnover” will be clear depending on whether the dense area is detected in the play of the boundary depending on whether it occurs before or after the “Lineout” Can be “Kick off” and “Kick counter”, which occur at short frame intervals before and after “Turnover” in GT, have similar ball positions, are difficult to distinguish, and frequently transition at short play intervals of “Penalty” → “Lineout” → “Turnover”. It is considered that because a ball is not detected in a certain “Lineout”, it is difficult to detect the ball position in the front and rear “Penalty” and “Turnover”, and it is difficult to distinguish the ball position.

In “Penalty”, from Fig. 8, Fig. 10, Fig. 12 in Recall, “Scrum” in “heatmap”, the part that was falsely detected as “Kick off” (FN) is reduced in “all”, but “Scrum”, It is easy to mistakenly detect as “Turnover”. This is probably because the distinction between “Scrum” and “Penalty” based on the presence or absence of a crowded area in the crowd and the distinction between “Kick off” and “Penalty” based on whether or not the ball detection position by the ball is on the line are clear. In addition, since the ball position is not detected in the portion where the transition is short at “Penalty” → “Lineout” → “Turnover” and the “Lineout” → “Scrum” in GT, the ball position is difficult to detect in the front and rear “Penalty” and “Turnover”, making it difficult to distinguish it is considered to be the body.

6 Conclusion

In this research, for six rugby plays, the variance of players in the “heatmap” feature value was clarified, and then the attribute of the subdivided area was added to the feature value for the average position of the players and the location of the dense area, and the dense play was performed. (“Scrum”, “Lineout”) and the other play and the play on the line (“Lineout”, “Kick off”) were determined, and the accuracy of the play was improved. With regard to the ball position, if we could accurately predict the players and the parts hidden behind the crowded area, we thought that it would be possible to add a new feature value and detect the offense and defense alternation as seen in “Turnover” more accurately.

References

1. Hochreiter, Sepp, and Jrgen Schmidhuber: LSTM can solve hard long time lag problems, In: Advances in neural information processing systems, 473-479(1997).
2. Simonyan, Karen, and Andrew Zisserman: Very deep convolutional networks for large-scale image recognition.
   In: arXiv1409.1556 (2014).
3. Kazunari Ouchi.: Development of a rugby video analysis system.
   In: Journal of the Institute of Electronics, Information and Communication Engineers B 100.12, 941-951(2017).
4. Von Gioi, R. G., Jakubowicz, J., Morel, J. M., & Randall, G.: LSD: A fast line segment detector with a false detection control.
   In: IEEE transactions on pattern analysis and machine intelligence, 32(4), 722-732(2008).
5. Hartley, Richard, and Andrew Zisserman: Multiple view geometry in computer vision.
   In: Cambridge university press(2003).
6. Beaton, Albert E., and JohnW. Tukey: The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data.
   In: Technometrics 16.2, 147-185(1974).
7. Cervone, D., D'Amour, A., Bornn, L., Goldsberry, K. :
   A multiresolution stochastic process model for predicting basketball possession outcomes.
   In: Journal of the American Statistical Association, 111(514), 585-599(2016).
8. Gudmundsson, Joachim, Thomas Wolle. :Football analysis using spatio-temporal tools.
   In: Computers, Environment and Urban Systems 47, 16-27(2014).