Pass Receiver Prediction in Soccer using Video and Players' Trajectories

Yutaro Honda[1] Rei Kawakami[2] Ryota Yoshihashi[2] Kenta Kato[3] Takeshi Naemura[1]
[1] The University of Tokyo [2] Tokyo Institute of Technology [3]Data Stadium Inc.
honda@hc.ic.i.u-tokyo.ac.jp reikawa@sc.e.titech.ac.jp

Abstract

In soccer, passing is one of the most fundamental actions for building tactics. Automatic prediction of the pass receiver can be useful in many situations, such as in player and team analysis and entertainment. In previous studies, the prediction is based on tracking data, in particular, time-series data of the two-dimensional positions of the players on the field, and little use has been made of video information such as the players' own posture and facial orientation. Thus, this paper aims to build a pass receiver prediction model that combines visual information with the trajectories of the players and the ball. We extract the features of the players' body movements from the video and the features of their movements on the field from the trajectories by using 3D convolutional networks and long short-term memory and learn the interactions between each player by using a transformer. Our study evaluation used wide-angle video and tracking data of 20 players, i.e.*, all players on the field excluding the goalkeepers. The results show that the prediction accuracy is greatly improved by using the video information.*

1. Introduction

Soccer is one of the most popular sports and has the largest global market size of any sport. Here, passing is the most frequent and dynamic part of soccer strategies, and it plays a big role in maneuvering the ball between teammates separated by short and long distances. If machines were able to predict the player who will receive the next pass or the location where a pass happens, it would have a wide range of applications; for example, they can be used to simulate what kind of pass will be selected in a particular situation or to measure the quality of a pass and its value on the field. This will facilitate analyses of players, team abilities, and tactical characteristics. Surprising or failed passes would be also clearer for non-experts.

Despite its usefulness, pass prediction remains one of the most difficult prediction tasks in soccer analysis. It requires high-level modeling of players' decision-making in temporal contexts. The quality and success of passes are also uncertain depending on the skill of the players or the situation. The most basic pass prediction models in the literature are location-based, where the players and the ball are represented by point trajectories, and hand-crafted features on them, such as velocity and acceleration, are utilized. These methods are not capable of handling visual information such as the players' posture or face orientation, which may help to improve the pass prediction. Some studies examined the usages of visual information in pass prediction but in a limited manner. For example, one study reported that simple video-based models do not match location-based models in terms of prediction accuracy [8]. Orientation-based models, which utilize vision-based pose estimation or geometric information such as the speed of the player , have been proposed [2, 26], but they still abstract rich visual information that the original video frames contain.

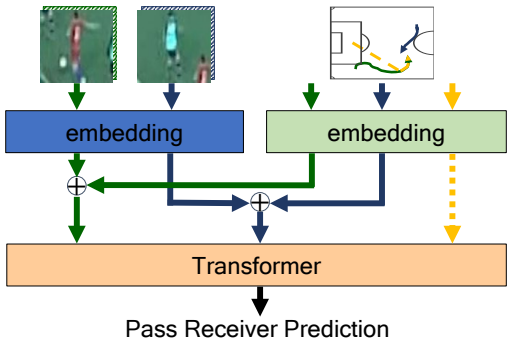


Figure 1. For each player, the visual and positional features are obtained separately from video and trajectory (green and blue arrows), then fused. The representations of the 20 players and the ball positional feature (yellow dotted arrow) are passed to a transformer encoder, which predicts the next ball receiver.

In this paper, we propose a method that learns to relate video frames with the location information of players and the ball for high-accuracy pass prediction. In our model, after the visual features and positional features of each player have been extracted using a 3D convolutional neural network (3D CNN) and long short-term memory (LSTM), it learns the relationship between players by using a transformer encoder to make predictions upon visual informa-

tion, location information, and the relationships between players. Fig. 1 illustrates the overview of our model.

To demonstrate the effectiveness of the proposed model, we obtained wide-angle match videos where 20 players on the field (excluding the goalkeepers) were always visible and tracking data that recorded the players' positions on the field at each point in time in the videos. In experiments with this data, our model improved top-1 accuracy by up to 13% compared with the model with only location information and had top-3 accuracy of over 90%. We also considered possible applications of our model and attempted two usages: detecting the timing of the changes in decision making and detecting high-level pass scene by prediction error.

2. Related work

Valuing plays is a major topic in analysis of soccer [3, 4, 13, 18, 19], and predicting the success rate of a pass or the probability of a pass being made may be used to design a metric. Power *et al*. used a linear regression model to calculate the success probability of a pass based on the distance and angle between each player, and calculate the reward of a pass based on whether a shot will occur within 10 seconds after receiving the pass [21]. Similarly, the Expected Possession Value (EPV), which is a continuous value from -1 to 1 that indicates how much a certain position on the field contributes to the goal, is a metric that uses the probability of success of a pass [10]. Furthermore, a pass prediction model suitable for EPV has been considered [9]. As shown above, pass direction and success rate play an important role in the evaluation of passes and plays, and better pass prediction models will produce better value indicators.

Traditional pass prediction methods use hand-crafted features, such as the distance between players defined in terms of the 2D coordinates of the players on the field, and use classical machine learning to predict the next player to receive a pass or the location where a pass will be made [7, 11, 17, 24–28]. Wei *et al*. predicted the pass receiver using a hidden conditional random field based on the speed, position, and direction of movement of 11 players in an arbitrary team; this was the first study in pass prediction to use tracking data [28]. Dauxais *et al*. used random forest [7] and 10 features related to the distance between players, such as the distance to the closest defender and their positions on the field immediately before a pass. Similarly, Li *et al*. extracted 54 features from the tracking data just before the occurrence of a pass, taking into account not only the distance and position but also the game situation and possible pass routes, and predicted the pass receiver by using LightGBM [17]. Hubacek *et al*. proposed a neural network for learning spatial features on the field by inputting hand-crafted features based on position and distance into the CNN [15]. Fernandez *et al*. created eight-feature-channels image-like tensors based on the position of the defender, the ball, and the goal at the moment of the pass, and trained the CNN to predict the location of the pass as a heat-map [9].

Most of the studies in which sports video is used as input are on comprehensive event/action recognition tasks [5, 12, 23, 29, 30]; only a few studies have treated individual action prediction tasks such as pass prediction. As for pass prediction using visual information, Felsen *et al*. used a simple fully convolution neural network to predict passes in water polo match videos, where the correct answer was the position in pixels of the player receiving the pass [8]. However, they reported that a prediction model using the player and ball position coordinates as input has higher accuracy. Arbues-Sanguesa *et al*. proposed a pass receiver prediction based on a probabilistic model combining the players' posture orientation (a kind of visual information), the positions of the enemy team players, and the distances between players [1]. Although the accuracy of the model was high, the posture direction of the players was estimated in advance with a deep neural network [2], and the prediction accuracy was affected by the accuracy of the estimation. As well, other information in the video image, such as the face orientation and posture state, was not used.

In this study, we attempted to input the video image itself in order to take into account the visual information included in the video image as much as possible and used a 3D CNN, which has been shown to be highly accurate in video recognition tasks. We also used the positional coordinates of each player to achieve highly accurate pass prediction.

3. Method

3.1. Overview

In this paper, we tackle *pass receiver prediction*, a task to predict the receiver of the next pass under the assumption that the next pass happens and succeeds, *i.e*., the pass is not interrupted or does not go out of play. The prediction can be seen as a k-way classification problem, where k means a number of the potential receivers. The task can be solved by assigning pass-receiving probability to each teammate, the conditional probability that a player receives a pass given that the pass happened. This can be denoted by

$$p(\mathbf{p}_i, t) = \frac{f(\mathbf{p}_i, t)}{\sum_{j \in [1,k], j \neq s} f(\mathbf{p}_j, t)}, \tag{1}$$

where $\mathbf{p}_i$ is i-th player, $\mathbf{p}_s$ is the pass sender, t is time frame, and f is the modeling function. Our dataset consists only of data from successful passes. This is because the failed passes are difficult to label who they were intended for. The input of our prediction model is a time series of frames of each player and their position on the field from a few seconds before the pass to the timing of the pass. The design philosophy is shown in Fig. 1.

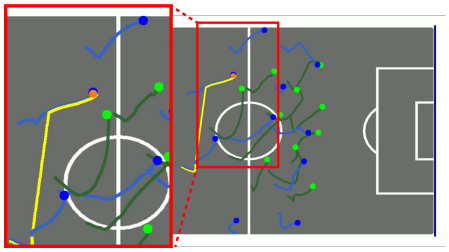


Figure 2. The green and blue lines indicate past trajectories of each player, and the green and blue dots are players' current position. The ball trajectory is illustrated in the yellow line and the orange dot is its current position.

We build our prediction model over wide-angle videos. There were several options for the type of video to be used, one of which was the broadcast video. However, not all players may be within the angle of view due to the camera work, and this will lead to data loss. Wide-angle videos, often captured by club teams for analytic purposes, give full tracks of the players and preferable for pass prediction. We also use their trajectories as illustrated in Fig. 2 that are obtained using a FIFA certified optical tracking system, called Tracab [6]. More details are described in Sec. 4.1.

3.2. Alignment between video and 2D trajectory

Our model assumes that the inputs are cropped player videos and players' 2D trajectories, as shown in Fig. 1. However, since the wide-angle videos do not have any annotations such as bounding boxes, we create our dataset by solving an alignment problem between players' position in the video and the 2D field coordinates obtained by the tracking system. Fig. 3 shows the overview of our process of alignment. First, we transform players' 2D positions in the field coordinate system to those in the video frames. Panoramic images of each match are generated and we use two homography matrices H_1 and H_2 for the coordinate transform. H_1 and H_2 are obtained by using video frames, panoramic images, and a few hand-picked 2D coordinates. Those transformed players' points still include positional errors because of camera distortion. To refine the positions, we detect players' positions in frames by You Only Look Once (YOLO) [22] version 5 [16] and align transformed player positions to those detected points in order to use tracked and labeled information in the original tracking data. However, there are two probable errors in detection of YOLOv5 and one issue for alignment: the first two are undetected players and detected unnecessary persons such as referees. The last one is uneven positional shifts, as depicted in the top center of Fig. 3.

For the issue of undetected players, we use a rigid point registration algorithm, iterative closest point (ICP) [31] and bring the two set of points as close as possible by a rigid transformation so that undetected positions in the video can be estimated. We refer to these points as pseudo-detected points. However, unnecessary points are also moved by ICP; thus, as depicted in Fig. 4, hungarian matching between the detected points by YOLOv5 and moved points by ICP is applied to filter required points. Finally, we use coherent point drift (CPD) [20], which calculates the mobility of each point individually, for non-rigid alignment between transformed points from tracking data and detected points by YOLOv5 with pseudo-detected points. By this process, the correct positions in the video coordinate and the size of bounding boxes of each player are obtained. More details are provided in **the supplementary material**.

3.3. Pass prediction

Our prediction model is designed to effectively combine soccer match videos and positional data on each player and the ball during the match. We consider the necessary elements for pass prediction are three-fold: the players' body movements, the players' positional movements on the field, and the interactions between the players. Thus, the prediction accuracy must be improved by selecting and combining suitable architectures for each element. The overview of our model is shown in Fig. 5.

Body motion embedding We use T_v video frames showing the 20 players of both teams except the goalkeepers before the exact timing of a pass as input to the feature extractor of the players' body movements. Since the area of a player in a wide-angle video is extremely small, if the entire video images were used as input, some players' information might be lost in the process of feature extraction. Therefore, we use cropped frames of each of the 20 players in the video obtained by our alignments process and extract each player's feature independently. In addition, it is desirable to use temporal context information to compensate for the low image resolution of the inputs; thus, we used a 3D CNN to extract features by convolution in the temporal dimension. We use a part of 3D resnet [14] and the weights of the model for feature extraction of each player were shared.

Trajectory embedding The input is 2D positions in the field coordinate system of the 20 players and the ball during T_t frames before the exact timing of the pass. A one-layer LSTM is used to extract the trajectory features for players' positional movements. In order to emphasize the movement just before the pass, which has a larger impact on the player's pass course selection, only the output of the last hidden layer of the LSTM is used, and the balance between the contribution of the past information and the information just before the pass is considered by using the forgetting mechanism of the LSTM. We extract the features for each player and the ball separately, whereas the weights of the LSTM are shared for all of them.

Learning interaction These trajectory and body move-

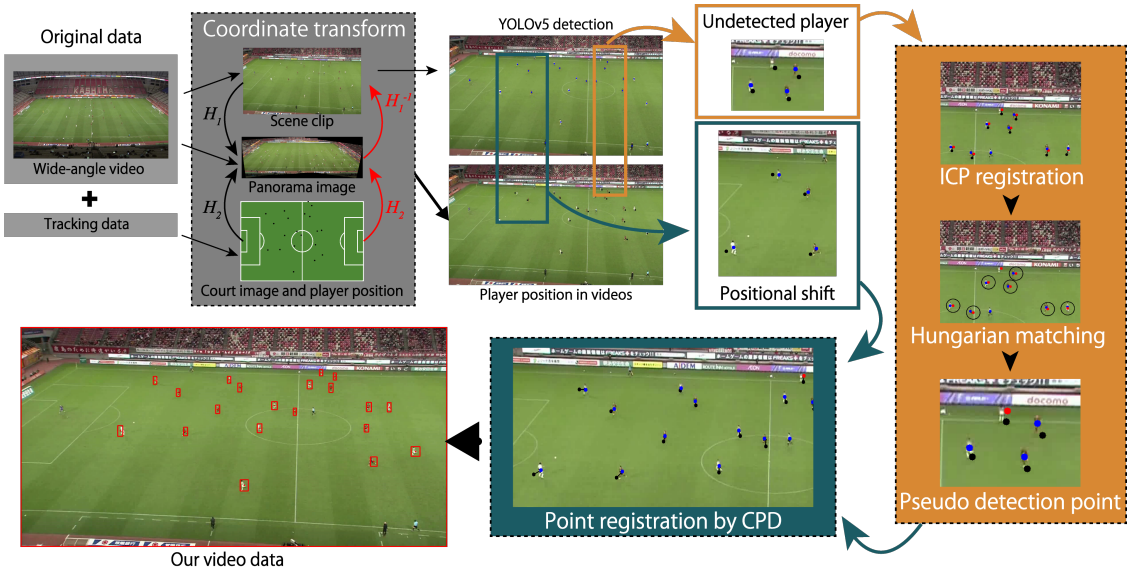


Figure 3. Overview of our processes for alignment between video and tracking data.

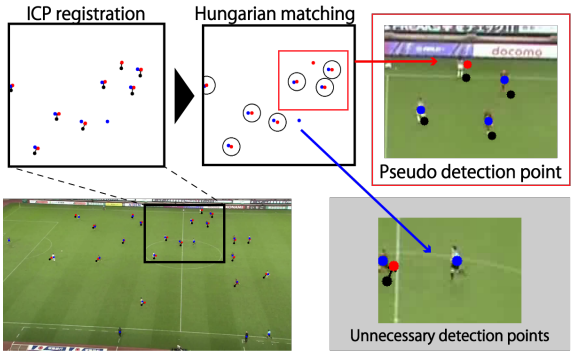


Figure 4. Addition of pseudo-detected points by ICP (red squares) and removal of unwanted points using Hungarian matching (gray squares). Blue points are obtained by YOLOv5.

ment features extracted for each player are summed, and the 20 features and the positional features of the ball are used as input to learn the interaction between the players. The relationship between players is modeled as a complete graph, because each players is directly or indirectly influenced by all the other players. We expect that the graph should model various perspectives (the relationship between teammates, opponents, surrounding players, etc.) through learning with a transformer. We also expect that the multi-head attention would enable the learning of multiple graphs that models multiple perspectives (more offensive/defensive, etc.). Namely, self attention can be understood as the process of learning complete graphs when the weight is the strength of the connections between nodes, and multi-head attention can be understood as generating multiple complete graphs because the attention weights are generated multiple times. We use a transformer encoder with 21 features (players and ball) as input and a residual connection between input and output in order to focus on learning interactions. The features are in the order of the passer, the 9 potential receivers, the 10 opposing players, and the ball, and we apply the position encoding just to indicate the order of input. While the model does not care the orders within potential receivers or opponent players, we aligned the receiver features and ground-truth receiver labels in the same place during training.

Finally, the probability of receiving a pass is predicted by applying the fully connected layers to each of the 9 players' features. The pass receiver is predicted by the output probability values from softmax operation.

4. Experiment

4.1. Setting

Dataset The dataset we used consisted of wide-angle videos in which 20 players (excluding the goalkeepers) appear at all times during a match, and tracking data showing the positions of all players including the goalkeeper, on the field at each point in time. The video and the tracking data were recorded in professional soccer league matches, and a total of 25 matches were obtained from three stadiums. The resolution of the match videos was 1920×1080. Since the sampling rate of the videos is 30Hz and the position coordinates is 25Hz, we resampled coordinates to 30Hz to synchronize them. For memory saving and faster training, we down-sampled only the videos to 15 Hz. The positional coordinates recorded were tracked by a dedicated optical tracking system and manually corrected. The range of the coordinate system was $(0, 0) \leq (x, y) \leq (5250, 3400)$. Although the coordinates of the ball were not recorded, the

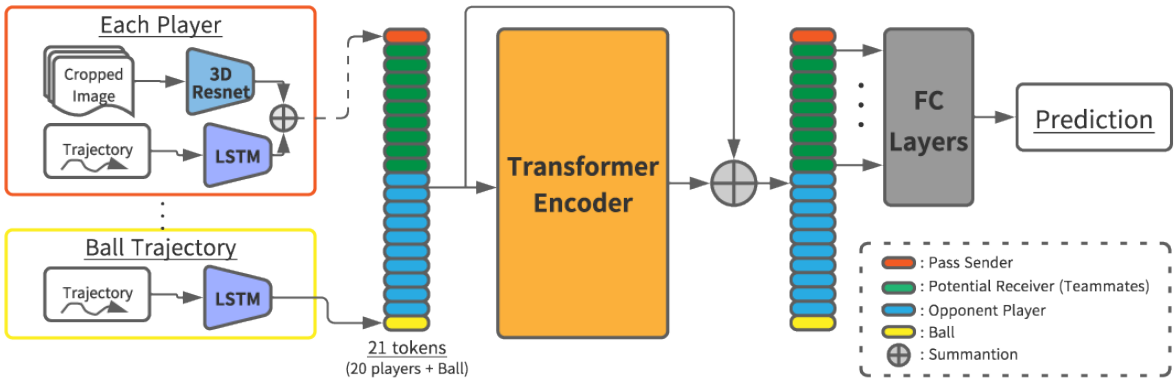


Figure 5. Overview of our network. For each player's representation, a video with T_v frames and a trajectory with T_t frames are processed by 3D Resnet [14] and a one-layer LSTM followed by feature fusing. A transformer encoder with residual connections updates the players' features while aggregating their interaction based on 20 players' tokens and the information. The ball-receive probabilities are predicted by a simple fully connected layer using the 9 teammates' features output by the transformer.

tracking data included the timing and location of events such as passes and shots; thus, the data for the ball were obtained by linear interpolation on the basis of the location and time of ball play. The coordinates of the players and the ball were normalized to $(-1, -1) \leq (x, y) \leq (1, 1)$, and the coordinates were horizontally flipped when necessary so that the attack direction of the team to which the passer belongs was aligned among the scenes.

Only successful pass scenes were used. Each scene has a length ranging from 1.0 to 5.0 seconds, and the pass occurs at the end of the scene. The total number of scenes was 15,586, of which 10,911 scenes were used for training, 1,559 scenes for validation, and 3,116 scenes for testing.

Implementation details We extracted body movement features from cropped clips consisting of 15 frames (1 second each) that were prepared for each player, and each frame was resized to 100×100 resolution. We used feature maps of intermediate layers and embeded them in 64 dimensions. For the trajectory feature extraction, we used 150 frames (5 seconds) showing the trajectories of each player and the ball as input in order to consider a longer time context. For the scenes shorter than 150 frames, we interpolated the trajectory data by zero-padding. The hidden layer had 64 dimensions. Note that the ball trajectory feature was used alone because there was no corresponding image feature. These features of the 20 players and the ball feature were passed to the transformer encoder and the outputs were fused with inputs by residual connection. The transformer had 4 layers and 4 heads; these settings were empirically determined. Finally, we passed the features that correspond to the potential receivers to the fully connected layers and converted into a pass-receive probability by using the softmax function.

Besides, we trained three other methods to compare their accuracy: our model without body movement features (only-trajectory), an existing method based on the position just before the pass [15], and a rule-based method that treats the closest teammate as the receiver. For the only-trajectory model, we used raw 25Hz trajectories instead of the resampled 30Hz trajectories. We used a batch size of 24 for training the proposed model and a batch size of 32 for training the trajectory-only model. For both models, the ADAM optimizer with hyper-parameters $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 10^{-8}$ was used and the learning rate was 0.0001. We used cross-entropy loss for training, and we used top-k accuracy for the test. To avoid overfitting, we stopped the models early and chose the models that had the top-1 accuracy on the validation data.

4.2. Quantitative evaluation

The experimental results are shown in Tab. 1. Our method reached the highest prediction accuracy for all top-k metric. Observing the increments of top-1 accuracy, taking the time series of positions into account is effective to improve the prediction. Also, including video information contributes to accurate prediction, and we can observe the largest impact for the improvement.

Next, we analyzed the difference of our proposed method and trajectory-only model in order to clarify the contribution of using videos. First, we paid attention to the relationship between prediction accuracy and pass distance. To reduce the risk of losing the ball, a player tends to pass it to the teammates close to them. Thus, the number of short passes is greater than that of long passes, making it difficult to predict long passes that rarely happen. However, intuitively, a player's kicking motion is an important key to estimating whether the player will pass on a short course or a long course. We analyzed the number of top-1 successful pass predictions and how far away the receiver was from the sender in that scene. As shown in the Tab. 2, the farther

Method	Accuracy (%)		
	top-1	**top-3**	**top-5**
Nearest	30.4	67.2	83.0
CNN ([15])	39.0	78.0	91.6
Ours (trajectory)	49.0	84.9	95.0
Ours (trajectory+RGB)	**62.5**	**92.3**	**97.5**

Table 1. Top-1/-3/-5 accuracy of the models.

Method	The number of successful predictions (scenes)								
	1st	**2nd**	**3rd**	**4th**	**5th**	**6th**	**7th**	**8th**	**9th**
Ours (trajectory)	641	378	248	123	60	43	14	13	8
Ours (trajectory+RGB)	**732**	**473**	**335**	**178**	**92**	**64**	**26**	**20**	**28**
Total number of scenes	1034	730	519	306	184	132	77	59	75
trajectory+RGB / trajectory	1.14	1.25	1.35	1.45	**1.53**	**1.49**	**1.86**	**1.54**	**3.5**

Table 2. The numbers of successful predictions when the true receiver is the n-th nearest neighbor from the sender.

away the receiver is, the fewer passes there are; this can leads to learning difficulties. Although the the number of successful pass predictions of both models were decreasing significantly, our model successfully predicted the receiver who was the seventh, eighth, or ninth nearest neighbor of the sender in about twice as many scenes as the trajectory-only model did. We consider the visual information helped to predict a long pass course.

Finally, the accuracy of the proposed model was compared by varying the video input length from 15 frames (1 second) to 75 frames (5 seconds). The model parameters were fixed for this analysis as the proposed model can handle such change of input dimension. The results are shown in Tab. 3. The overall trend is that the accuracy decreases as the amount of temporal input increases. The decrease is especially large for the top-1 accuracy. The reason for this may be that the parameter size is not appropriate for the video length. Since the parameter size is the same for an input volume that is up to five times larger, it may not be possible to extract information using the entire video, so it is desirable to increase the parameter size. However, it is known through experiments that learning becomes unstable or overfitting occurs extremely quickly if the parameters are made too large; thus, it is necessary to search for the correct parameter size and structure.

Top-k accuracy (%)	Video length (frames)				
	15	30	45	60	75
Top1	**61.10**	59.89	59.25	*57.59*	57.98
Top3	**91.52**	91.17	90.31	*89.65*	90.00
Top5	**97.47**	97.03	97.15	96.88	*96.53*

Table 3. The top-k accuracy v.s. video length. The accuracy is the average of three test models.

4.3. Qualitative evaluation

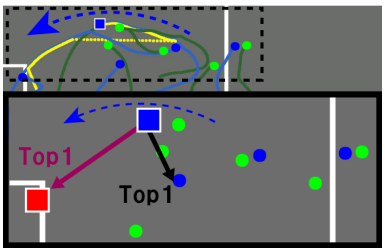
On the basis of the scores of the quantitative evaluation, we qualitatively analyzed the predictions of the existing methods. First, we analyze the difference between the **trajectory-only model** and the **CNN model** [15] to grasp their tendency. Then we visualize the difference between the predictions of the **trajectory-only model** and the **proposed model (trajectory+RGB)**.

Regarding the effect of the time-series positional information, we found that the prediction is more successful in scenes where the player passes while moving, such as during dribbling. This is because players are likely to pass the ball to the player who is ahead of them in the direction of their trajectory, as can be seen in Fig. 6a. In addition, the trajectory-only model tends to predict the pass course more toward the opponent's goal because there are many such scenes when the player's team is on the attacking side. However, this is also related to prediction failure, because the pass prediction in a different direction from the trajectory direction often fails, as shown in Fig. 6b. In particular, in the case of a direct pass or a pass when the passer is stationary and there are several teammates around him who are about the same distance away, it is difficult to predict who the receiver will be based on the the trajectory information alone. For those scenes, visual information is helpful.

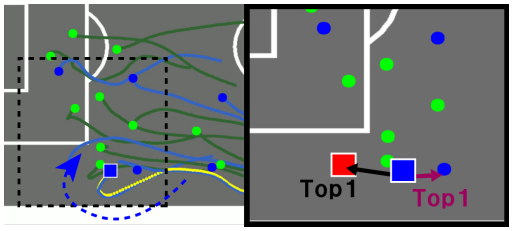
Next, we compare our model to the trajectory-only model. Fig. 7 shows another example of scenes which trajectory-only model cannot handle well. Although the sender did not move largely before receiving and passing the ball directly, our model (trajectory+RGB) predicted the right pass course (the red arrow) by taking into account the kicking form. Our model is also capable of predicting the correct direction of a pass even when the pass is in a different direction from the trajectory. Fig. 8 is the same scene to the one in Fig. 6b. The passer's kick motion corrects for the adverse effect of the trajectory information. Similarly, the prediction accuracy of long passes was improved by taking the kicking form into account, especially the swinging of the feet.

As an example of the usefulness of the video information of the receiver, we found that the prediction of a pass to a position where it was difficult to make a pass, such as a through pass to a place where there were many opponent players, was successful when the receiver's motion of receiving the ball was considered. Fig. 9 shows such a case.

However, all the models tended to fail in circumstances where it is difficult to control the ball and the ball tends to go where the player does not intend, such as in heading and floating passes. For the same reason, in scenes of long passes for a side change, there were many cases where the direction of the pass was correct, but the predicted player and the actual receiver were different.



(a) The trajectory helps to predict the correct pass course when the sender passes while moving.



(b) The trajectory-based model becomes erroneous when predicting a pass in a different direction from the player's trajectory.

Figure 6. The bird's eye view images of a soccer field and trajectories. The area in the black rectangle is a zoomed view of the area in the black dotted rectangle. The blue and red squares indicate the sender and the receiver (both of them in the blue team). The wine-red arrows show the top-1 predictions by the **trajectory-only model**, and the black arrows show the top-1 predictions by the **CNN model** [15]. The blue dotted arrows represent the outlines of the senders' trajectories.

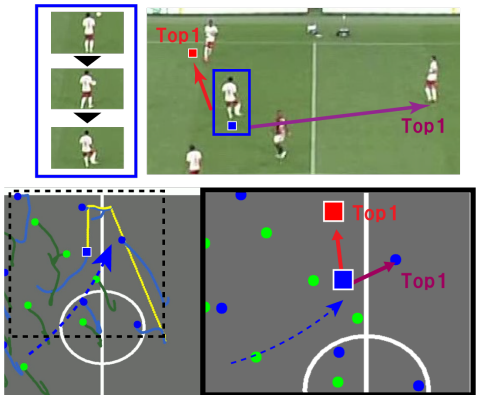


Figure 7. A scene of a direct pass. The red arrow represents the top-1 prediction by **our model**. The sender's kicking movement is observed from video frames in the left blue box. Inputting video frames corrected the prediction by trajectory-only model.

4.4. Limitations

Our model has several limitations. The first is that the successful predictions even with our model are the passes to the nearest neighbors. In this study, we used only the successful passes in our data, and successful passes tend to be in short distance. In other words, our prediction model trained on this data has a bias toward predicting short distances. The passes that are failed to be predicted were likely to be in longer distances. By including the failure passes in the learning, we should be able to create a more realistic pass prediction model. Second, the model does not take the game theory into account or game situations such as the score difference and remaining time. Such information may provide a clue to improve the prediction accuracy in scenes where the prediction probabilities of the top-2 and top-3 are almost the same, namely, when the model is not sure which course to predict.

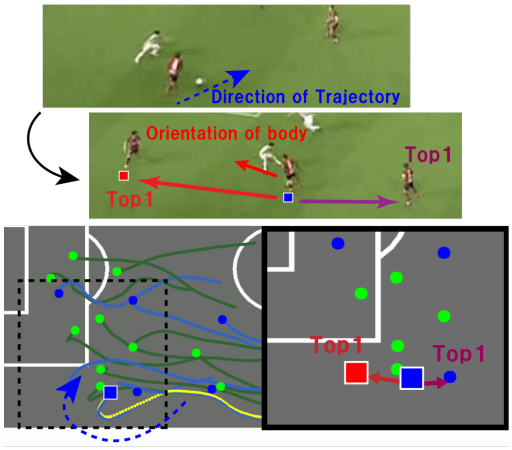


Figure 8. A scene when the player stopped to change the direction right before the pass. The red arrow shows prediction by **our model**, and the wine red shows that by trajectory-only model. Body orientation becomes a key feature for prediction.

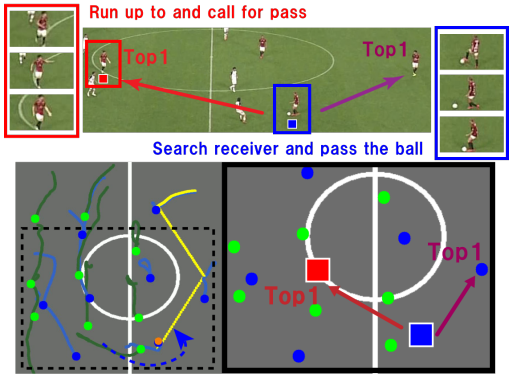


Figure 9. An example of scenes which the sender's action (blue box on the right) and the receiver's action (red box on the left) contributed to accurate prediction. The red and wine-red arrows are predictions of ours and trajectory-only model.

5. Possible Application

In this section, we describe a possible application of our model. We show that our model can analyze the play based on the probability change during a time frame before a pass

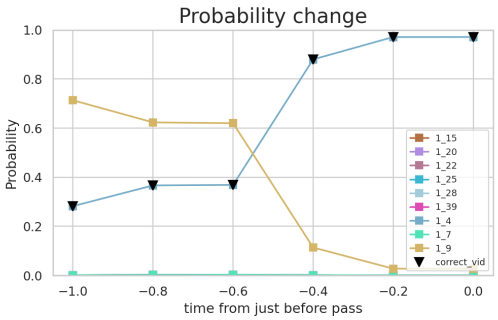


Figure 10. A scene when two specific players' probability values are changing largely. It seems that the sender decided to change the pass receiver at -0.5 seconds before the pass happens.

happens.

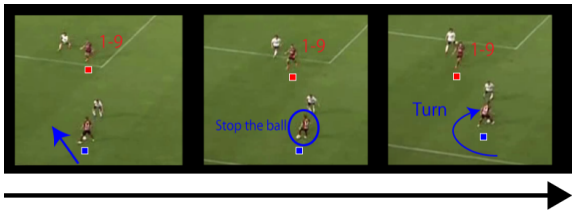
We used our model to observe the probability during one second before a pass happened and detected the moments when the sender changed his decision. We input videos and trajectories from 0.0, 0.2, 0.4, 0.6, 0.8, and 1.0 seconds before the pass into our model, and recorded the probability values output at each of these timings. Fig. 10 is an example of the players' probability change. Each line shows the change in the probability value of potential receivers at each time. The number in the legend is the player identification ID, which corresponds to the number in the video in the analysis described below. The black downward triangle indicates the player who actually receives the pass. The horizontal axis shows the timestamps before the pass: -1.0 indicates 1.0 seconds before the pass and 0.0 indicates the exact timing of the pass. In this scene, ID1-9 (yellow) has a high probability of receiving the ball at the beginning, and the probability for ID1-4 (blue, the correct receiver) gradually increases from the 0.5 seconds before to the time the pass happened. Since the probability of receiving the ball for the other players is low, it is assumed that the sender changed the target of the pass from ID1-9 to ID1-4 at around -0.5 seconds.

The actual images of this scene are shown in Fig. 11. The blue squares represent the sender, and the red squares represent the ID1-9 and 1-4 with high probability values. This scene starts with two players approaching the sender to receive a pass as in Fig. 11a. As can be seen in Fig. 11b, the sender turns his body into the side space to pass the ball as the player 1-9 comes in. This movement and body direction would have increased the probability value for the player 1-9.

Next, we checked the actual video of the subsequent increase in the probability value of the player 1-4. In the video, we can observe that the sender stops the ball, turns to another direction, and finally chooses to pass the ball to the player 1-4 as in Fig. 11c. This behavior is considered to be a factor that resulted in a higher probability for the player 1-4.

(a) The first frame of the scene: -1.0 time points at Fig. 10.

(b) Video frames of Fig. 10 from -1.0 to -0.6 seconds. The sender (blue) turns his body to face the player 1-9 who comes in.



(c) Video frames of Fig. 10 from -0.6 to -0.2 seconds before the pass occurred. You can see the sender stop the ball once and make a turn.

(d) The last frame of this scene.

Figure 11. The actual video frames of the target scene shown in Fig. 10

Thus, we can see that the change in the probability values correctly reflects the change in the players' pass choices. By analyzing the change in the probability values, it is possible to analyze what pass choices the players were trying to make.

6. Conclusion

We described a new pass prediction model that is based on the visual and location information of the players and the relationships among the players. We have presented specialized modules for each of them and proposed an architecture that allows the video images to be input directly. This makes it possible to consider the player's visual information without prior processing such as pose estimation. In addition, we described the comprehensive procedure for creating a dataset that combines video and tracking data.

References

[1] A. Arbues-Sanguesa, A. Martin, J. Fernandez, Coloma Ballester, and Gloria Haro. Using Player's Body-Orientation to Model Pass Feasibility in Soccer. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 3875–3884. IEEE, jun 2020. 2

[2] A. Arbues-Sanguesa, A. Martin, J. Fernandez, C. Rodriguez, G. Haro, and C. Ballester. Always Look On The Bright Side Of The Field: Merging Pose And Contextual Data To Estimate Orientation Of Soccer Players. In *2020 IEEE International Conference on Image Processing (ICIP)*, volume 2020-Octob, pages 1506–1510. IEEE, oct 2020. 1, 2

[3] Alina Bialkowski, Patrick Lucey, Peter Carr, Yisong Yue, Sridha Sridharan, and Iain Matthews. Large-Scale Analysis of Soccer Matches Using Spatiotemporal Tracking Data. *Proceedings - IEEE International Conference on Data Mining, ICDM*, 2015-Janua(January):725–730, 2014. 2

[4] Sanjay Chawla, Joël Estephan, Joachim Gudmundsson, and Michael Horton. Classification of Passes in Football Matches Using Spatiotemporal Data. *ACM Transactions on Spatial Algorithms and Systems*, 3(2):1–30, aug 2017. 2

[5] Anthony Cioppa, Adrien Deliege, Silvio Giancola, Bernard Ghanem, Marc Van Droogenbroeck, Rikke Gade, and Thomas B Moeslund. A Context-Aware Loss Function for Action Spotting in Soccer Videos. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13123–13133. IEEE, jun 2020. 2

[6] ChyronHego Corporation. Tracab gen5 earns fifa certification, 2020. Last accessed 25 January 2022. 3

[7] Yann Dauxais and Clément Gautrais. Predicting Pass Receiver in Football Using Distance Based Features. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 11330 LNAI, pages 145–151. 2019. 2

[8] Panna Felsen, Pulkit Agrawal, and Jitendra Malik. What will Happen Next? Forecasting Player Moves in Sports Videos. In *2017 IEEE International Conference on Computer Vision (ICCV)*, volume 2017-Octob, pages 3362–3371. IEEE, oct 2017. 1, 2

[9] Javier Fernández and Luke Bornn. SoccerMap: A Deep Learning Architecture for Visually-Interpretable Analysis in Soccer. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 12461 LNAI, pages 491–506, 2021. 2

[10] Javier Fernández, Luke Bornn, and Dan Cervone. Decomposing the Immeasurable Sport: A deep learning expected possession value framework for soccer. *MIT Sloan Sports Analytics Conference*, pages 1–18, 2019. 2

[11] Philippe Fournier-Viger, Tianbiao Liu, and Jerry Chun-Wei Lin. Football Pass Prediction Using Player Locations. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 11330 LNAI, pages 152–158. 2019. 2

[12] Kirill Gavrilyuk, Ryan Sanford, Mehrsan Javan, and Cees G. M. Snoek. Actor-Transformers for Group Activity Recognition. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 836–845. IEEE, jun 2020. 2

[13] Laszlo Gyarmati and Xavier Anguera. Automatic Extraction of the Passing Strategies of Soccer Teams. aug 2015. 2

[14] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6546–6555, 2018. 3, 5

[15] Ondřej Hubáček, Gustav Šourek, and Filip Železný. Deep Learning from Spatial Relations for Soccer Pass Prediction. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 11330 LNAI, pages 159–166. 2019. 2, 5, 6, 7

[16] Glenn Jocher, Alex Stoken, Jirka Borovec, NanoCode012, Ayush Chaurasia, TaoXie, Liu Changyu, Abhiram V, Laughing, tkianai, yxNONG, Adam Hogan, lorenzomammana, AlexWang1900, Jan Hajek, Laurentiu Diaconu, Marc, Yonghye Kwon, oleg, wanghaoyang0106, Yann Defretin, Aditya Lohia, ml5ah, Ben Milanko, Benjamin Fineran, Daniel Khromov, Ding Yiwei, Doug, Durgesh, and Francisco Ingham. ultralytics/yolov5: v5.0 - YOLOv5-P6 1280 models, AWS, Supervise.ly and YouTube integrations, Apr. 2021. 3

[17] Heng Li and Zhiying Zhang. Predicting the Receivers of Football Passes. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 11330 LNAI, pages 167–177. 2019. 2

[18] Daniel Link, Steffen Lang, and Philipp Seidenschwarz. Real Time Quantification of Dangerousity in Football Using Spatiotemporal Tracking Data. *PLOS ONE*, 11(12):e0168768, dec 2016. 2

[19] Tatsuya Mimura and Yohei Nakada. Quantification of pass plays based on geometric features of formations in team sports. In *ACM International Conference Proceeding Series*, pages 306–313, 2019. 2

[20] Andriy Myronenko and Xubo Song. Point set registration: Coherent point drifts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(12), 2010. 3

[21] Paul Power, Hector Ruiz, Xinyu Wei, and Patrick Lucey. "Not All Passes Are Created Equal:" Objectively Measuring the Risk and Reward of Passes in Soccer from Tracking Data. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1605–1613, New York, NY, USA, aug 2017. ACM. 2

[22] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 779–788, 2016. 3

[23] Ryan Sanford, Siavash Gorji, Luiz G Hafemann, Bahareh Pourbabaee, and Mehrsan Javan. Group Activity Detection from Trajectory and Video Data in Soccer. In *2020 IEEE/CVF Conference on Computer Vision and Pattern*

Recognition Workshops (CVPRW), pages 3932–3940. IEEE, jun 2020. 2

[24] Yusuke Sano and Yohei Nakada. Improving Prediction of Pass Receivable Players in Basketball. In *Proceedings of the Tenth International Symposium on Information and Communication Technology - SoICT 2019*, pages 328–335, New York, New York, USA, 2019. ACM Press. 2

[25] Samriddha Sanyal. Who will receive the ball? Predicting pass recipient in soccer videos. *Journal of Visual Communication and Image Representation*, 78(May):103190, jul 2021. 2

[26] William Spearman, Austin Basye, Greg Dick, Ryan Hotovy, and Paul Pop. Physics-Based Modeling of Pass Probabilities in Soccer. In *MIT Sloan Sports Analytics Conference, Boston (USA),*, number March, pages 1–14, 2017. 1, 2

[27] Vincent Vercruyssen, Luc De Raedt, and Jesse Davis. Qualitative spatial reasoning for soccer pass prediction. *CEUR Workshop Proceedings*, 1842, 2016. 2

[28] Xinyu Wei, Patrick Lucey, Stephen Vidas, Stuart Morgan, and Sridha Sridharan. Forecasting events using an augmented hidden conditional random field. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 9006, 2015. 2

[29] Jianchao Wu, Limin Wang, Li Wang, Jie Guo, and Gangshan Wu. Learning Actor Relation Graphs for Group Activity Recognition. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9956–9966. IEEE, jun 2019. 2

[30] Lifang Wu, Zhou Yang, Qi Wang, Meng Jian, Boxuan Zhao, Junchi Yan, and Chang Wen Chen. Fusing motion patterns and key visual information for semantic event recognition in basketball videos. *Neurocomputing*, 413:217–229, nov 2020. 2

[31] Zhengyou Zhang. Iterative point matching for registration of free-form curves and surfaces. *International Journal of Computer Vision*, 13(2):119–152, oct 1994. 3