End-to-End High-Risk Tackle Detection System for Rugby

Naoki Nonaka[1] Ryo Fujihira[1] Monami Nishio[1] Hidetaka Murakami[2]
Takuya Tajima[3] Mutsuo Yamada[4] Akira Maeda[5,6] Jun Seita[1]
[1]Advanced Data Science Project, RIKEN Information R&D and Strategy Headquarters
[2] Murakami Surgical Hospital [3] Faculty of Medicine, University of Miyazaki
[4] Faculty of Health and Sport Sciences, Ryutsu Keizai University
[5] Hakata Knee & Sports Clinic [6] Faculty of Human Health, Kurume University

Abstract

Reducing risk of severe injury such as concussion is a high priority for any contact sports. In rugby, Head Injury Assessment (HIA) protocol has been introduced to identify and protect players showing symptoms of concussion and having potential risk of concussion. However, on-field decisions by officials are sometimes difficult and subjective, and HIA is affordable only for elite leagues since it requires medical specialists. To make rugby matches more safe, we aim to develop a system to detect high-risk tackles, potential triggers of concussion, based on deep learning models. Our system takes rugby match video, then first identifies frame with tackle, subsequently detects location of tackle and estimate pose of the ball carrier and the tackler, and finally evaluate the risk of tackle using posture pair of players. Among the model combinations we have examined, the best performance was achieved with the combination of ResNet (2+1)D as tackle frame selection model, RetinaNet as tackle detection model and CenterTrack as pose estimation model. Evaluation using test data, a set of short clips from broadcasted rugby match videos, showed our system was able to detect 50% of high-risk tackles without any human intervention. This result opens a path for automated systems to detect high-risk events, leading to less expensive and more objective monitoring not only for rugby but also for any contact sports.

1. Introduction

Rugby Union (or 'rugby') is a fast-paced collision sport, with a high incidence rate of various injuries. The most common injury in Rugby World Cup (RWC) was concussion ($n = 24$, 13.9% of all injuries in RWC 2015 [8]; and $n = 22$, 15.4% of all injuries in RWC 2019 [6]). A meta-analysis revealed that overall incidence of match-play concussion in men's rugby was 4.73 per $1,000$ player match hours [9]. The occurrence of concussion not only causes the loss of training time but also raises the risk of harmful aftereffect later in life [21]. Such a negative aspect of rugby had long been ignored [24]. International Rugby Board (renamed to 'World Rugby') declared that the risk management of concussion became a key strategy in rugby [27].

To take an appropriate treatment and minimize the risk of sequelae, it is important to keep athletes out of participation when there is any suspicion of concussion [20]. In 2015, World Rugby introduced a matchday concussion management protocol, referred to as Head Injury Assessment (HIA) [7]. HIA is consist of three processes aimed to identify players with suspected concussion and accurately diagnosed by official match day doctors (details are written in [7]). However, the on-field decision of suspected concussion is subjective, competing with a time constraint and athletes eager to play. There exist a certain percentage of overlooking with or without intention. In addition, due to the high cost of deploying medical specialists, only the matches such as elite level and international games can afford official match day doctors monitoring.

Taking advantage of recent progress in deep learning, we aim to develop a system that detects high-risk events automatically from rugby match videos. Since 76% of concussion in rugby is caused by tackle [35], we focus our detection target to tackles with the risk of causing concussions. In the previous study about automated tackle risk assessment [19], tackle scenes were identified manually in advance, making it difficult to apply their system to raw videos. In contrast, we aim to develop a detection system for high-risk tackle from rugby match video without human intervention.

This paper is organized as follows. First, Section 2 describes the related studies, and Section 3 describes overview of our high-risk tackle detection system and details of data and models used for our system. Then, in Section 4, we explain the evaluation metric used to evaluate the overall system and the results are presented in Section 5. Finally,

discussion and conclusions are given in Section 6.

2. Related Work

With the success in domains such as image processing, natural language processing and speech recognition, the application of deep learning is increasing in the field of sports. The applications of deep learning is making sports games and training more efficient, more logical and even less dangerous. For example, automatic action spotting has been conducted in various sports like soccer [3, 13, 31], baseball [25], and ice hockey [37]. Likewise, pose estimation has been utilized in swimming [4,38,39], running [4,33], table tennis [1, 16]. The main purposes of these studies are to analyze games and to quantify player's performance. Also, there are some studies aimed to improve player's safety, such as prediction of injuries in baseball pitchers [26] or automatic injury detection in soccer [22].

However, in the field of rugby, the application of deep learning hardly progress. The reason for this difficulty is that rugby games have a lot of contact plays, which create occlusions where particular actions or players are hidden by other players. Occlusions make it difficult to estimate players' behavior from video frames. So far, only one study has conducted with deep learning framework in rugby [19]. They used YOLO [28] to detect a ball carrier and a tackler and manually extract high risk tackles with their own definition. Though, in this study, easy-to-detect 1-on-1 tackles were manually extracted from videos as a dataset in advance. Utilizing deep learning in rugby game is still a frontier of research. Other studies of rugby are based on manually extracted features or rule-based calculation of measurement data, which are conducted to classify dangerous tackles [10,36] or to detect contact plays [12,15]. These approaches are still time-consuming due to manual data extraction or complicated data preprocessing.

3. High-risk Tackle Detection System

In this study, we propose a system to detect high-risk tackles from rugby match videos. As shown in Fig.1a, the proposed system consists of four models: tackle frame selection model, tackle detection model, pose estimation model, and tackle risk classification model. The system takes a video as input, classifies frames by tackle frame selection model, combines tackle detection model and pose estimation model to detect tackles and finally classifies risk of given tackles. A high-risk tackle is defined as a tackle that leads to a Head Injury Assessment (HIA) in the official record. In the following, we describe the details of each model that composes our system and datasets used to train and evaluate each model.

3.1. Data

As a source for training and evaluation of our system, we used rugby match videos of Japanese elite league and corresponding official match records. A total of 360 videos from three seasons of the Top League, an elite rugby league in Japan, from the 2016 to 2018 seasons were used to build datasets. The videos used were broadcasted on TV, with frame sizes ranging from 854×480 to 1920×1080 and fps of 24 or 25. The HIA records were obtained from the official match records available on the official page of the Top League[1].

Following the procedures shown in Fig. 1b, we prepared videos used in our experiment. First, we selected the match video with HIA based on the official match record. Subsequently, we identified 226 frames which contain event resulted in HIA by manually checking the videos. Among the identified frames, we selected 230 frames in which HIAs were caused by tackle as high-risk tackle frame[2]. Of 360 videos, 87 videos contained at least one high-risk tackle frame. We randomly split these 87 videos into training and test set with $9 : 1$ ratio. The training and test set videos were used to build the datasets necessary for training and evaluating the models composing the system.

3.1.1 Dataset for training of tackle frame selection model

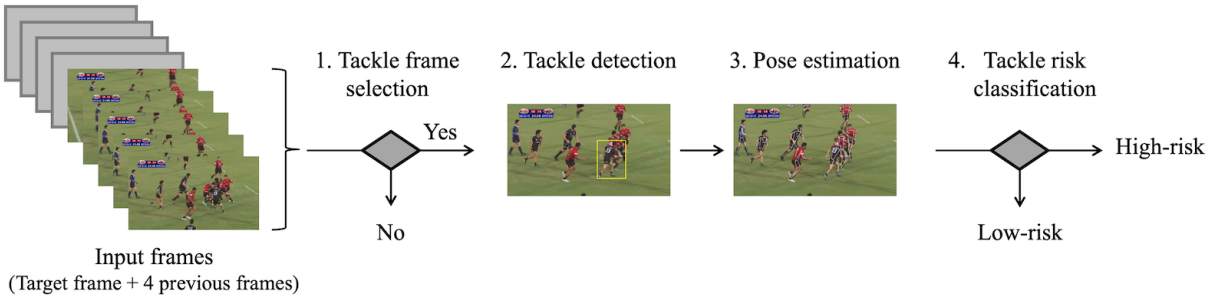
By using the videos assigned to training set, we constructed a dataset to train a frame selection model to determine whether a given video clip contains a tackle or not. As shown in Fig. 1c, we randomly selected 100 video clips of 2 seconds in length each from 78 videos assigned to the training set, for a total of $7,800$ videos. Subsequently, we manually checked each video clip and labeled whether the final frame of video clip contains tackle or not. As a result, there were 199 video clips with and $7,601$ video clips without tackle in final frame. These $7,800$ video clips were randomly divided into $8 : 2$ and used as the dataset for training and evaluation of the tackle frame selection model.

3.1.2 Dataset for training of tackle detection model

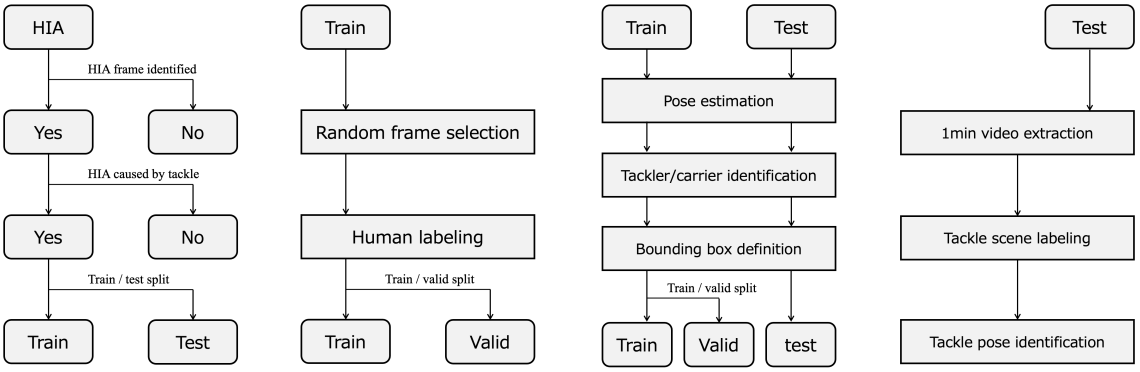
A dataset for tackle detection model was constructed as follows. The procedure is shown in Fig. 1d. We used videos with high-risk tackle frame, and from each video, we selected low-risk tackle frame, frames with tackle not leading to HIA. We identified the one to three times of low-risk tackle frame per high-risk tackle in each video, resulting

[1] https://www.top-league.jp/

[2] Some frames were extracted from highlight which shows same event from diffrent angle and with diffent magnification. The unique count of high-risk tackle was 149.



(a) Overview of high-risk tackle detection system.



(b) Selection of videos with high-risk tackles.

(c) Dataset preparation procedure for tackle frame selection model.

(d) Dataset preparation procedure for tackle detection model.

(e) Dataset preparation procedure for overall system evaluation.

Figure 1. Overview of the system to detect high-risk tackles and preparation procedure for datasets used to train models. We define a high-risk tackle as a tackle that has led to a HIA in the official record. (a) Our high-risk tackle detection system is composed of four models, namely tackle frame selection model, tackle detection model, pose estimation model and tackle risk classification model. For each rugby match video, we first split video into chunks of 5 frames and select chunks containing tackle in final frame using tackle frame selection model. Subsequently, we apply the tackle detection model and pose estimation model to the final frame of chunks selected as tackle. By combining both results, we use extracted posture of players inside detected tackle location and apply model from [23] to classify risk of tackles. (b) Using video with HIA identified by official match record, we first identify frame of HIA event, subsequently select HIA caused by tackles and split selected video into training and test set with $9:1$ ratio. (c) To build a dataset for training tackle frame selection model, we randomly selected 100 frames from videos split into training set, manually labeled each frame and split frames into training and validation set. (d) To build dataset for training tackle detection model, we first applied CenterTrack [41] to all videos, secondly identified high-risk tackle frames, thirdly identified same number of low-risk tackle, subsequently identified tackler and carrier on each frame, defined bounding box using detected pose of tackler and carrier, and finally split videos contained in training set into training and validation set. (e) To build dataset for evaluation of overall detection system, first we randomly extract 1 minute video from videos split into test set in (d), manually labeled all tackle frames and identified all tackle poses.

in total of 400 low-risk tackle frames. Subsequently, we applied pose estimation by CenterTrack [41] to these high-risk tackle and low-risk tackle frames, and identified tackler and ball carrier. After identifying tackler and ball carrier, we selected frames in which 5 or more key-points were successfully detected for both tackler and ball carrier. As a results, 155 out of 230 high-risk and 238 out of 400 low-risk tackle frames were selected respectively. From the coordinates of the extracted tackler and ball carrier postures, we defined rectangular area covering both posture as the tackle area, and labeled it as the detection target. Finally, we randomly split the frames from the training set video into training set and validation set frames, and used them to train the tackle detection model described in Fig. 1b.

3.1.3 Dataset for evaluation of overall system

To evaluate overall high-risk tackle detection system, we prepared video clips as shown in Fig. 1e. First, we extracted 65 video clips of one minutes length from test set videos. The 33 video clips were manually created by placing the high-risk and low-risk tackle frames identified in Fig. 1d at the center of video clip. To make video clip with tackle frame at the center, we cut out 30 seconds of video clip before and after the identified tackle frames. The other 32 video clips were extracted by randomly selecting frames and cutting out one minute length video clips from selected

frames. Subsequently, all frames contained in 65 extracted video clips were manually checked and given a label to indicate whether they contain tackle or not. As a result, total number of 65 video clips, with 12 high-risk tackle video clips, were obtained.

3.2. Components of the system

Our system is composed of four models, namely tackle frame selection model to classify given frame contains target or not, tackle detection model to find location of tackle in given frame, pose estimation model to obtain posture of players in given frame and tackle classification model to classify given postures are a high-risk tackle or not. Our system takes frames from rugby matches as an input and outputs binary labels indicating high-risk tackles exist in corresponding frame, and is sequentially applied to evaluate whole video.

3.2.1 Tackle frame selection model

To exclude frame without tackle we classify frames by frame selection model. The frame selection model takes set of target frame to classify and four frames before target frame as input and determines whether given target frame contains a tackle or not. In this study, we tested three video classification models, ResNet Mixed Convolution, ResNet (2+1)D, and ResNet3D [34], which are used as frame selection models. All three models used were pre-trained with Kinetics-400 dataset [14] and fine-tuned with dataset described in section 3.1.1. Since the output of original model was 400 dimensions in accordance with the Kinetics-400 dataset, we added two fully connected layers to create a binary classification model in which each frame was a tackle scene or not. The number of frames input to the model was set to five, and the size of each frame was resized to 224×224 to match the size of the Kinetic-400 dataset. The inverse of the ratio of tackle labels to non-tackle labels was used as the weight of the tackle class to mitigate class-imbalance problems during training. For each model, we selected the optimal learning rate, batch size, and optimizer through preliminary experiments. For data augmentation during training, we randomly extracted 5 frames from the original 2-second (50-frame) video at every epoch and fed them to the model, as well as randomly changing the color.

3.2.2 Tackle detection model

After selecting tackle frames, to detect location of tackle in each frame, we apply a tackle detection model. We used object detection models which detect rectangular regions containing the postures of ball carrier and tackler in the input image. We tested three models, namely DETR [2], RetinaNet [17], and YOLOv3 [29]. All models were pre-trained using COCO dataset [18] and fine-tuned using prepared tackle location data described in section 3.1.2. As a data augmention random flips were performed during the training. For each model, we selected optimal learning rate and batch size based on the result of the preliminary experiment. The model with the best performance in the validation set was selected as the final model.

3.2.3 Pose estimation model

Following the detection of tackle, we apply pose estimation model to frames selected by tackle detection model. As a pose estimation model, two models, namely HRNet [32] and CenterTrack [41] were used. For both models, we used publicly available models pretrained with COCO dataset [18] with no additional training. To extract posture of players related to tackle, we first apply pose estimation model to extract posture of all players in the given frame. Subsequently, we automatically extract tackle related players, namely ball carrier and tackler by assuming player is related to tackle, if player's part of torso (region of body surrounded by both shoulders and both sides of the waist) is located inside tackle region given by tackle detection model. In summary, pose estimation model takes frames selected by the tackle selection model and tackle location given by tackle detection model as an input and outputs postures of players related to tackle.

3.2.4 Tackle risk classification model

After extracting posture of tackle related players, we classify whether tackle in given frame is high-risk or not by applying tackle classification model. In this study, we classify risk of tackles by using tackle related players' posture pair. To classify the risk of tackle, we use Naive Bayes model from [23] to classify given pair of postures is high-risk or not. If three or more postures are related to tackle, we take all combination of pairs and evaluate each pair by Naive Bayes model.

4. Evaluation Metric

To evaluate the detection performance of events in time series data in balance with false positives, we defined a utility function based on [30]. We defined a positive example as the frame 1.5 seconds before and after the high-risk tackle. If even one tackle that was judged as high-risk tackle in the interval corresponding to a positive example, the model was treated to have detected a high-risk tackle. All frames other than those defined as positive examples were defined as negative examples. As shown in Fig. 2, each frame was assigned a score of $+1$ if the prediction result was true positive, -1 if it was false negative, -0.1 if it was false positive, and 0 if it was true negative. The score of all frames in each video was calculated and the total exhibition of frames in all

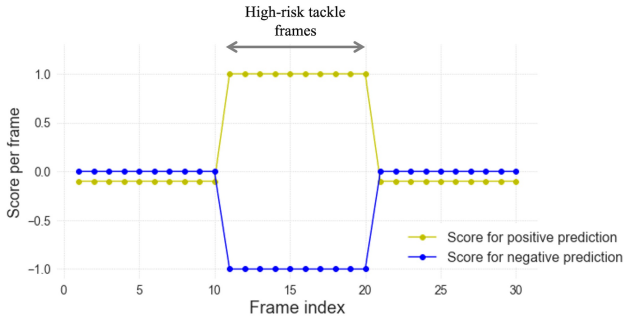


Figure 2. Evaluation metric for detection of high-risk tackles from rugby match video. To evaluate time series of frames, we use a modified evaluation metric based on [30]. We evaluate each frame based on a utility function which gives a score of $+1$ to true positive frames, -1 to false negative frames, -0.1 for false positive frames and 0 for true negative frames. Subsequently, we sum up scores for each frame existing in all videos in the test set to obtain U_{total}. We also calculate score U_{max} which corresponds to the case when all frames are predicted positive and U_{neg} which corresponds to the case when all frames are predicted negative.

65 videos was calculated as U_{total}. In addition, we calculated U_{max} as the score when all frames were correctly classified, and U_{neg} as the score when all frames were predicted to be negative. Then, U_{total}, U_{max} and U_{neg} were normalized by the following equation to obtain the final score.

$$U_{score} = \frac{U_{total} - U_{neg}}{U_{max} - U_{neg}} \tag{1}$$

5. Result

We evaluated our high-risk tackle detection system using test set videos shown in Fig. 1b. We first evaluated the frame selection model, tackle detection model and pose estimation model independently. Subsequently, we evaluated the overall performance of our model.

5.1. Tackle frame selection model

First, we examined the performance of the frame selection model for classifying tackle frames. We evaluated the performance of frame selection model using video clips described in Fig. 1e. The results are shown in Table 1. ResNet Mixed Convolution showed better performance compared to ResNet 3D in terms of macro F1 and recall, showing F1 score of 0.56 and recall of 0.2. In the rugby game, there are various tackle patterns and the video used in this study was broadcast video, so the angles of shooting were also various, which made it difficult to improve the classification performance. In addition, when the class threshold for binary classification was reduced from 0.5 to 0.1, the precision decreased to about 0.2, but the recall improved to about 0.4. This result suggests that if false positives can be tolerated by the tackle detector downstream of the system, the overall system performance could be improved by changing the threshold value.

5.2. Tackle detection model

Next, we evaluated the performance of tackle detection model. Table 2 shows the performance on the frames from test set. The evaluation metric was calculated by comparing the intersection over union (IoU) between true bounding box in the test set and the predicted bounding box in each condition. The evaluation metrics were calculated in the following three patterns.

1. Top confidence bbox IoU: Draws only the bounding box with the highest confidence among the bounding boxes detected in each frame, and calculates the IoU with the true bbox as the evaluation value.

2. Average bbox IoU: Sets a confidence threshold, calculates the IoU with the true bounding box for all bounding boxes above the threshold, and uses the average value as the final value.

3. Best bbox IoU: Set the confidence threshold, calculate the IoU between the bounding box with the highest confidence and the true bounding box among the bounding boxes above the threshold, and use the average value as the final value.

In addition to the metrics described above, we calculated the ratio of images with successful detection of tackles in the test set, where we assume detection was successful if the IoU between any of the detected bboxes and the true bounding box is greater than 0. As a result, RetinaNet and DETR had similar losses and equal numbers of successful detection of tackles, but YOLOv3 performed worse than the two models. In the following analysis, only two models, DETR and RetinaNet, were included in the analysis.

Next, we applied RetinaNet and DETR to rugby match videos to qualitatively evaluate their performance of tackle detection. Fig. 3 shows a typical examples of successful detection and some failure patterns. As can be seen from Fig. 3b and 3c, compared to RetinaNet, DETR was likely to detect more tackles, resulting in more false positives with less false negatives. In both models, the typical patterns of false positives were scrum (shown in Fig. 3d), one player holding the ball, and two players making contact without the ball.

5.3. Pose estimation model

Subsequently, we evaluated the qualitative performance of the pose estimation model. We applied two pose estimation models, namely CenterTrack and HRNet to the test set video described in Fig. 1b respectively. The typical examples of result of applying pose estimation models are shown

Table 1. Result of frame selection model on test set video from Fig. 1e. All three models were pretrained with Kinetics-400 dataset [14], and fine-tuned with training set shown in Fig. 1c. We compare three classifiers and the case without applying classifier (No classifier; assuming all frames as tackle frame). For all three classifiers, threshold of positive and negative class was set to 0.5. ResNet (2+1)D and ResNet Mixed Convolution showed better performance compared to ResNet 3D in terms of macro F1 and recall.

Frame selection model	Macro F1	Recall	Precision
No classifier	0.114	1.	0.136
ResNet Mixed Convolution	0.564	0.199	0.312
ResNet (2+1)D	0.565	0.21	0.301
ResNet 3D	0.534	0.127	0.275

Table 2. Result of tackle detection model applied to frames in the test set shown in Fig. 1d. All three models were trained with COCO dataset [18] and fine-tuned using training and validation set shown in Fig. 1d. We evaluated models with three metrics, namely top confidence bbox IoU, average bbox IoU and best bbox IoU, together with ratio of detection. RetinaNet and DETR showed similar performance on all four metrics, while performance of YOLOv3 was worse in all metrics.

	Top confidence bbox IoU	Average bbox IoU	Best bbox IoU	ratio of detection
DETR	0.647	0.646	0.679	0.939 (31/33)
RetinaNet	0.655	0.577	0.655	0.939 (31/33)
YOLOv3	0.277	0.277	0.277	0.364 (12/33)

in Fig. 4. Fig. 4a show pose estimation results of zoom out image, in which pose estimation by CenterTrack was successful while HRNet failed detect any postures. When applied to zoom in image, as shown in Fig. 4b, both models were successful in estimating posture of players. As shown in Fig. 4c, both models failed to detect players when occlusion such as scrum existed in the frame.

5.4. Overall system evaluation

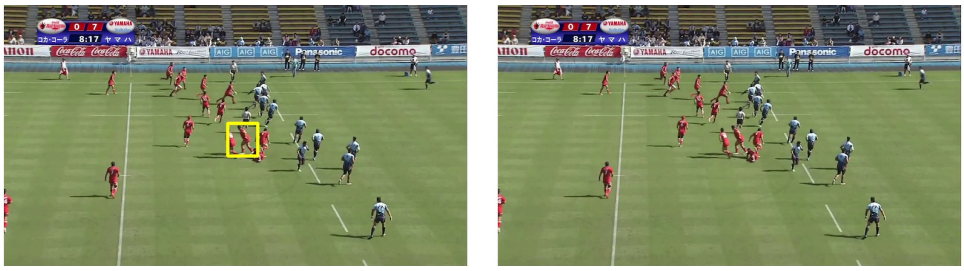
Finally, we evaluated the overall performance of the high-risk tackle detection system. We tested all combination of four frame selection models (setting to select all frames together with three frame classification models), two tackle detection models (DETR and RetinaNet) and two pose estimation models (HRNet and CenterTrack). In addition to the above settings, we evaluated the performance of the system when frame selection model was replaced by human labels. The results of applying the system to test set video described in Fig. 1e is shown in Table 3. Among the models tested, the best performance was obtained when ResNet (2+1)D was used as the frame selection model, RetinaNet as the tackle detection model, and CenterTrack as the pose estimation model. In terms of individual models, the pose estimation model using CenterTrack outperformed the one using HRNet. For the tackle detection model, the scores of RetinaNet and DETR were comparable. Considering the case without frame selection model and the frame selection based on human labels, DETR performed better when the number of false positives in the frame selection model was small, while RetinaNet performed better when the number of false positives was large. The tackle detection model used in this study tends to have more false positives in DETR than in RetinaNet, indicating the importance of balancing the false positives with those in frame selection model. Since recall tends to be low in the frame selection model used in this study, improving this point may improve the overall performance.

6. Discussion

In this study, we developed a system to detect high-risk tackles from rugby match videos. We defined high-risk tackle as a tackle that led to HIA recorded in the official match record. Our system was composed of four models, namely frame selection model to select frames with tackles, tackle detection model to detect location of tackle in the frame, pose estimation model to estimate posture of players and tackle classification model to classify high-risk and low-risk tackles. Among the combinations of models we tested, the best performance was achieved with combining ResNet (2+1)D as a frame selection model, RetinaNet as a tackle detection model and CenterTrack as a pose estimation model. Overall evaluation of the system showed that our high-risk tackle detection system was able to detect 50% of high-risk tackle in held out videos.

Our system can be applied to rugby match video directly without a need of human intervention. Thus, we believe our system can contribute to the development of automated high-risk tackle detection applied to rugby match. However, our system has several limitations, such as requirements of multiple deep neural network models resulting in slow processing speed, failure of pose estimations when players are occluded and false positive classification of high risk tackles. The processing speed may improved by using smaller

(a) Example of an image in which both DETR (left) and RetinaNet (right) were successful in detecting a tackle.



(b) Example of an image in which a false positive occurs in DETR (left) but not in RetinaNet (right).

(c) Example of an image in which correctly detected in DETR (left) but not in RetinaNet (right).

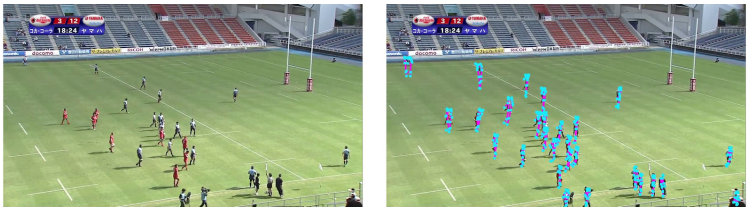
(d) Example of an image in which both DETR (left) and RetinaNet (right) failed in detecting a tackle.

Figure 3. Example of tackle detection with DETR and RetinaNet applied to rugby match. False positives were more frequent in DETR (left column) than in RetinaNet (right column), while false negatives tended to occur more frequently in RetinaNet. False positives were seen in both DETR and RetinaNet in situations where people occluded, such as in a scrum.

networks, such as MobileNet [11] as backbone. Occlusion problems maybe improved by applying models using information of privious frames [40, 41] or applying models which estimate depth from monocular camera image [5]. As for the mitigation of false positive classification result of tackle postures, replacement of the rule-based selection of tackle related players may work. The rule-based selection of players tends to increase the number of players selected, resulting in increase of likelihood of getting at least one high-risk tackles, leading to evaluate input frame as a frame with high-risk tackle. One potential way to alleviate this problem is to add ball detection model to the system, as implemented in [19]. In the future study, we would like to improve the overall performance of our system to apply in the real-world.

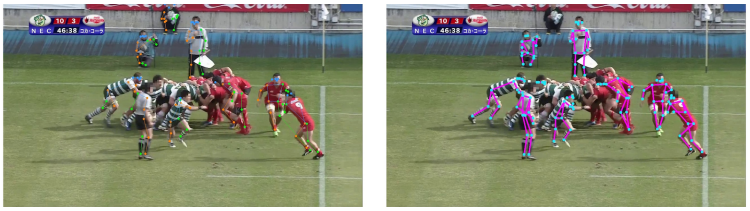
Acknowledgement

This work was supported by MEXT "Innovation Platform for Society 5.0" Program Grant Number JPMXP0518071489.



(a) Example of pose estimation with HRNet (left column) and CenterTrack (right column)

(b) Example of an zoom in image, both HRNet (left) and CenterTrack (right) succeeded.



(c) Example of an image with occlusion, both model failed with occluded players.

Figure 4. Example of pose estimation with HRNet (left column) and CenterTrack (right column) applied to rugby match. (a) At zoom out image, pose estimation by HRNet failed, while CenterTrack succeeded. (b) At zoom in image, both HRNet and CenterTrack was successful in estimating pose. Qualitatively performance of CenterTrack was better compared to HRNet. (c) In the case of human occlusion, such as scrum, both methods often failed to estimate the pose.

Table 3. Result of evaluation on 1 minute video shown in Fig. 1e. We evaluate all combination of three frame selection models, two tackle detection models and two pose estimation models. We also evaluate a case when frame selection model was replaced by human labels and a case without frame selection. Score was highest with combination of ResNet (2+1)D, RetinaNet and CenterTrack with score of 0.2807 and recall of 0.5.

Frame selection model	Tackle detection model	Pose estimation model	Score	Recall
Human labels	RetinaNet	HRNet	0.3449	0.583
		CenterTrack	0.4905	0.833
	DETR	HRNet	0.2249	0.417
		CenterTrack	0.5397	0.917
No selection	RetinaNet	HRNet	0.2312	0.583
		CenterTrack	0.2759	1.000
	DETR	HRNet	0.2204	0.583
		CenterTrack	0.2224	1.000
ResNet Mixed Convolution	RetinaNet	HRNet	0.1837	0.333
		CenterTrack	0.0793	0.167
	DETR	HRNet	0.1825	0.333
		CenterTrack	0.1680	0.333
ResNet 2+1D	RetinaNet	HRNet	0.0840	0.167
		CenterTrack	0.2807	0.500
	DETR	HRNet	0.000	0.000
		CenterTrack	0.2719	0.500
ResNet 3D	RetinaNet	HRNet	0.0867	0.167
		CenterTrack	0.0400	0.083
	DETR	HRNet	0.0866	0.167
		CenterTrack	0.0820	0.167

References

[1] Lewis Bridgeman, Marco Volino, Jean-Yves Guillemaut, and Adrian Hilton. Multi-person 3d pose estimation and tracking in sports. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0, 2019. 2

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 4

[3] Anthony Cioppa, Adrien Deliege, Floriane Magera, Silvio Giancola, Olivier Barnich, Bernard Ghanem, and Marc Van Droogenbroeck. Camera calibration and player localization in soccernet-v2 and investigation of their representations for action spotting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4537–4546, 2021. 2

[4] Moritz Einfalt and Rainer Lienhart. Decoupling video and human motion: towards practical event detection in athlete recordings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 892–893, 2020. 2

[5] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2002–2011, 2018. 7

[6] CW Fuller, Aileen Taylor, Marc Douglas, and Martin Raftery. Rugby world cup 2019 injury surveillance study. *South African Journal of Sports Medicine*, 32(1), 2020. 1

[7] Colin W Fuller, Gordon W Fuller, Simon PT Kemp, and Martin Raftery. Evaluation of world rugby's concussion management process: results from rugby world cup 2015. *British journal of sports medicine*, 51(1):64–69, 2017. 1

[8] Colin W Fuller, Aileen Taylor, Simon PT Kemp, and Martin Raftery. Rugby world cup 2015: world rugby injury surveillance study. *British journal of sports medicine*, 51(1):51–57, 2017. 1

[9] Andrew J Gardner, Grant L Iverson, W Huw Williams, Stephanie Baker, and Peter Stanwell. A systematic review and meta-analysis of concussion in rugby union. *Sports medicine*, 44(12):1717–1731, 2014. 1

[10] Michaela Hopkinson, Garetha Nicholson, Dana Weaving, Shariefa Hendricks, Annaf Fitzpatrick, Adamf Naylor, Colinf Robertson, Clivea Beggs, and Bena Jones. Rugby league ball carrier injuries: The relative importance of tackle characteristics during the european super league. *European journal of sport science*, pages 1–10, 2020. 2

[11] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1314–1324, 2019. 7

[12] Billy T Hulin, Tim J Gabbett, Rich D Johnston, and David G Jenkins. Wearable microtechnology can accurately identify collision events during professional rugby league matchplay. *Journal of Science and Medicine in Sport*, 20(7):638–642, 2017. 2

[13] Haohao Jiang, Yao Lu, and Jing Xue. Automatic soccer video event detection based on a deep neural network combined cnn and rnn. In *2016 IEEE 28th International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 490–494. IEEE, 2016. 2

[14] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 4, 6

[15] Daniel Kelly, Garrett F Coughlan, Brian S Green, and Brian Caulfield. Automatic detection of collisions in elite level rugby union using a wearable sensing device. *Sports Engineering*, 15(2):81–92, 2012. 2

[16] Kaustubh Milind Kulkarni and Sucheth Shenoy. Table tennis stroke recognition using two-dimensional human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4576–4584, 2021. 2

[17] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 4

[18] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 4, 6

[19] Zubair Martin, Sharief Hendricks, and Amir Patel. Automated tackle injury risk assessment in contact-based sports-a rugby union example. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4594–4603, 2021. 1, 2, 7

[20] Paul McCrory, Willem Meeuwisse, Jiří Dvorak, Mark Aubry, Julian Bailes, Steven Broglio, Robert C Cantu, David Cassidy, Ruben J Echemendia, Rudy J Castellani, et al. Consensus statement on concussion in sport—the 5th international conference on concussion in sport held in berlin, october 2016. *British journal of sports medicine*, 51(11):838–847, 2017. 1

[21] Marc P Morissette, Heather J Prior, Robert B Tate, John Wade, and Jeff RS Leiter. Associations between concussion and risk of diagnosis of psychological and neurological disorders: a retrospective population-based cohort study. *Family medicine and community health*, 8(3), 2020. 1

[22] OV Ramana Murthy and Roland Goecke. Injury mechanism classification in soccer videos. In *ICCV Workshops*, pages 774–779, 2015. 2

[23] Monami Nishio, Naoki Nonaka, Ryo Fujihira, Hidetaka Murakami, Takuya Tajima, Mutsuo Yamada, Akira Maeda, and Jun Seita. Objective detection of high-risk tackle in rugby by combination of pose estimation and machine learning. *in preparation*. 3, 4

[24] Jon S Patricios and Simon Kemp. Chronic traumatic encephalopathy: Rugby's call for clarity, data and leadership in the concussion debate, 2014. 1

[25] AJ Piergiovanni and Michael S Ryoo. Fine-grained activity recognition in baseball videos. In *Proceedings of the ieee conference on computer vision and pattern recognition workshops*, pages 1740–1748, 2018. 2

[26] AJ Piergiovanni and Michael S Ryoo. Early detection of injuries in mlb pitchers from video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 2

[27] Martin Raftery. Concussion and chronic traumatic encephalopathy: International rugby board's response, 2014. 1

[28] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. 2

[29] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018. 4

[30] Matthew A Reyna, Chris Josef, Salman Seyedi, Russell Jeter, Supreeth P Shashikumar, M Brandon Westover, Ashish Sharma, Shamim Nemati, and Gari D Clifford. Early prediction of sepsis from clinical data: the physionet/computing in cardiology challenge 2019. In *2019 Computing in Cardiology (CinC)*, pages Page–1. IEEE, 2019. 4, 5

[31] Ryan Sanford, Siavash Gorji, Luiz G Hafemann, Bahareh Pourbabaee, and Mehrsan Javan. Group activity detection from trajectory and video data in soccer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 898–899, 2020. 2

[32] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5693–5703, 2019. 4

[33] Kazunari Takeichi, Masaru Ichikawa, Ryota Shinayama, and Takehiro Tagawa. A mobile application for running form analysis based on pose estimation technique. In *2018 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pages 1–4. IEEE, 2018. 2

[34] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018. 4

[35] Ross Tucker, Martin Raftery, Gordon Ward Fuller, Ben Hester, Simon Kemp, and Matthew J Cross. A video analysis of head injuries satisfying the criteria for a head injury assessment in professional rugby union: a prospective cohort study. *British journal of sports medicine*, 51(15):1147–1151, 2017. 1

[36] Ross Tucker, Martin Raftery, Simon Kemp, James Brown, Gordon Fuller, Ben Hester, Matthew Cross, and Ken Quarrie. Risk factors for head injury events in professional rugby union: a video analysis of 464 head injury events to inform proposed injury prevention strategies. *British journal of sports medicine*, 51(15):1152–1157, 2017. 2

[37] Kanav Vats, Mehrnaz Fani, Pascale Walters, David A Clausi, and John Zelek. Event detection in coarsely annotated sports videos via parallel multi-receptive field 1d convolutions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 882–883, 2020. 2

[38] Dan Zecha, Moritz Einfalt, Christian Eggert, and Rainer Lienhart. Kinematic pose rectification for performance analysis and retrieval in sports. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1791–1799, 2018. 2

[39] Dan Zecha, Moritz Einfalt, and Rainer Lienhart. Refining joint locations for human pose tracking in sports videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0, 2019. 2

[40] Yifu Zhang, Chunyu Wang, Xinggang Wang, Wenjun Zeng, and Wenyu Liu. Fairmot: On the fairness of detection and re-identification in multiple object tracking. *International Journal of Computer Vision*, 129(11):3069–3087, 2021. 7

[41] Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. Tracking objects as points. *ECCV*, 2020. 3, 4, 7