Risky Tackle Detection from American Football Practice Videos using 3D Convolutional Networks

Nasik Muhammad Nafi[1], Scott Dietrich[2], and William Hsu[1]

[1] Kansas State University, Manhattan, KS 66502, USA
{nnafi,bhsu}@ksu.edu
[2] Barry University, Miami Shores, FL 33161, USA
sdietrich@barry.edu

Abstract. In this paper, we introduce the problem of risky tackle detection from American football practice videos and propose a 3-stage Convolutional Neural Network (CNN)-based pipeline to improve detection accuracy. At first, we propose an anomaly detection-based approach to temporally localize the tackle action. Spatial regions of interest are then identified using an object recognition model. Finally, 3D convolution is applied to classify risky and safe tackles based on spatiotemporal features. Our approach trades off between end-to-end action classification from untrimmed videos and precise localization of temporal anchors of an action. We conduct our experiment on a newly created data set that contains 178 annotated videos collected from seven different practice fields. We empirically demonstrate that our proposed method outperforms state-of-the-art video classification and anomaly detection approaches applied directly to untrimmed tackle videos.

Keywords: American Football, Head Injury, Risky Tackle Identification, Deep Learning, Sports Video Classification.

1 Introduction

In this work, we address the problem of simultaneous action detection and risk estimation as a classification task for videos. Such computer vision applications in sports span a gamut of static scene analysis to high frame rate videos covering entire practice sessions, and from brief training exercises to plays. The key rationale for visual analysis of sports videos is monitoring and then providing early warnings of potentially injurious practices. This would allow coaches to intervene to prevent injury and mitigate resulting risk and harm, whether physical, psychological, or financial, from improper tackling. Specifically, the Centers for Disease Control (CDC) estimates that between 1.6 and 3.8 million sports-related concussions (SRC) are reported annually with American football showing the highest proportion of head injuries or concussions among all sports [21] [5]. Research shows that in youth football, on an average, one player out of every 33 players may suffer a concussion during the season. Concussions occur at a rate

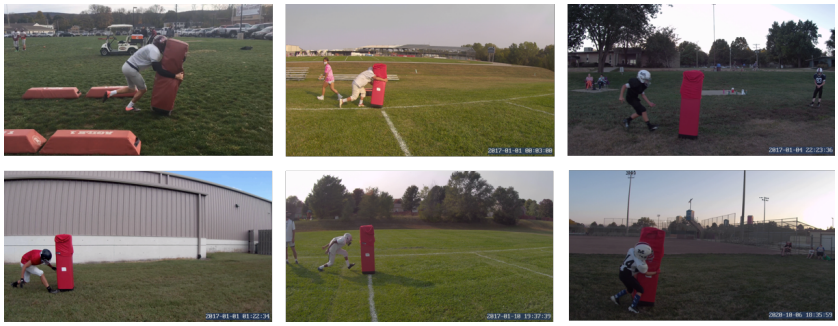

Fig. 1: Representative frames from videos collected at different practice fields.

of 9.9 per 10,000 athlete exposures; where each athlete exposure is considered one play either in practice or in a game [22]. In addition, head impact may cause brain injuries such as hemorrhage, hematoma, and edema. Annually, millions of dollars are spent to treat injured players and maintain reserved players [36]. Furthermore, this adversely affects the teams, both in competitive performance and reputation.

Two-thirds of all football-related head injuries occur during practice and one-third during games, 47% of all SRC occur as a result of head-to-head collisions [4]. Researchers have found that early exposure to American football may have a long-term neuropsychiatric and cognitive effects such as Chronic Traumatic Encephalopathy (CTE) due to repeated head impacts [29] [1]. Learning proper tackle form at an early age is an important developmental milestone for reducing unnecessary head impacts among youth football players [21] [23].

Identification and correction of improper-tackle techniques is a key step for establishing a safe playing environment. Coaches wanting to reduce the potential for player to player head impacts may choose to use blocking dummies when teaching the skill to young players. Practice tackles are filmed so that athletic trainers and coaches can identify dangerous postures and provide corrective feedback on player performance. However, these video assessments are carried out manually by human judges [28] [39] [19]. Manual processing of the videos to classify risky or safe tackle requires a substantial amount of effort and time from human assessors.

CNN-based architectures have been shown to be successful at extracting novel visual features directly from RGB images [13] [38]. Use of CNN in tandem with Long Short-Term Memory (LSTM) network and the introduction of 3D convolution have made a breakthrough in many video processing tasks such as activity recognition, event or action localization, anomaly detection [17] [8] [40] [16] [3] [37]. This influenced researchers to adopt deep learning-based computer vision approaches in sports analytics [10] [31]. However, the inherent differences in actions performed in different sports pose a different set of challenges.

Automatic detection of the risky forms of tackle solely based on videos can greatly improve a coach's ability to correct player behavior and reduce the likelihood that the players sustain head impacts. More importantly, this will help the players to find out the overall safety ratings of their tackles just after performing them rather than waiting for more than a week while the coaches analyze the videos. To the best of our knowledge, no prior research has attempted to classify tackles from videos of American football practice. In our work, we first exploit an anomaly detection mechanism to temporally segment the informative frames containing the tackle and then leverage an existing state-of-the-art object detection model to extract regions of interest from those frames. In last stage, a customized 3D ConvNet is used to classify risky and safe tackles from the spatiotemporally segmented frame sequence.

To summarize, our key contributions are as follows:

- We introduce the task of risky tackle detection directly from videos of American football practice with a tackle dummy.
- We present a set of 178 labeled American football tackle practice videos collected in the United States.
- We propose a framework for detecting risky tackles from practice videos using only video-level annotation.
- We conduct a comparative analysis of our proposed pipeline with state-of-the-art video classification and anomaly detection approaches for untrimmed video and evaluate the results in terms of precision, recall, and F1-score.

2 Related Work

Video Classification: One of the core tasks of video processing is video classification, commonly referred as activity recognition. In the last few decades, it was very common to use hand-crafted features for video representation. Spatiotemporal interest points (STIPs) [27], 3D variants of scale-invariant feature transform (SIFT-3D) [33], and histogram of oriented gradients (HOG-3D) [24], improved Dense Trajectories (iDT) [41] demonstrated promising results. Recent CNN-based approaches have already gained success over those hand-crafted features [17] [8] [40]. One common approach is to extract frame-level features using 2D convolution followed by Long Short-Term Memory (LSTM) cells to capture temporal dynamics from those frame-level features [8]. 3D ConvNet eliminates the need for LSTM blocks by extending 2D convolution into the temporal dimension, making it well-suited for spatiotemporal feature learning directly from video [40] [16].

Two-stream networks [35] [3] utilize both RGB frames and optical flow frames. Optical flow can capture apparent motion information invariant to appearance. The RGB and flow frames are fed into identical ConvNet to extract features and are fused at some particular stage. C3D [40], I3D [3], R(2+1)D have proved that a video classification network trained on a sufficiently large data set such as

Kinetics[18], Sports-1M [17] can be used to extract video features for completely different tasks from other domains.

A different approach, however, is to consider binary video classification as an anomaly detection problem. Most anomaly detection approaches use unsupervised or semi-supervised methods such as dictionary learning [45], topic modeling [15], histograms [6], or autoencoders [44] to learn the distribution of normal video, so it can distinguish the anomalies. Some recent approaches attempt to solve the problem with supervised learning using both normal and anomalous videos with video-level annotation [37].

Action Localization: Most techniques for action localization assume that untrimmed input videos are annotated with the temporal anchor of the action. They then treat the task as an iterated image classification task, where the system needs to classify each candidate window derived from running a temporal sliding window over the whole video [30] [10]. More recently, to reduce the number of candidate windows, temporal action proposals [9] [14] have been introduced. Buch et al. [2] presented a single-stream temporal action proposal (SST) to mitigate the issue of multiple passes over the same video frames. However, all of these approaches require manual annotations for atomic actions which is subjective, laborious, and time-consuming. Shou et al. [34] proposed a multi-stage CNN to solve the temporal localization problem. Although it relaxes the requirement of exact temporal annotation, its benefit is overshadowed by the complexity of multi-scale candidate segment generation and multiple network training.

Approaches that completely forgo action-level temporal annotation generally use a learning framework for multiple instance selection [25] [26] [37]. This allows localization of action or anomalous events by finding key instances in untrimmed videos. The video segments are considered instances, and the key instances are learned based on only video-level labels.

Object Detection: Object detection refers to identifying an object and its localization. Region-Based Convolutional Neural Networks (R-CNN) [11] have shown impressive results in object detection. The process includes a sequence of CNN-based feature extraction, object classification, and bounding box regression. Mask R-CNN generates a mask in pixel level of the object to segment it from the generated proposals [12]. Faster-RCNN uses a dedicated CNN-based Region Proposal Network (RPN) that drastically reduces the proposal generation time [32].

Injury Detection: Injury detection or prediction is a well-studied area in sports analytics. However, most research, especially for American football, are based on either physical and psychological statistics of the player [7] [20] or data collected from micro-sensor [19] [42] and manual investigation of incident videos [28] [39] [19]. Very recently, [31] successfully applied 3D convolution to early detection of injury in baseball pitchers using only videos. There is hardly any video-based work for American football that attempts to identify risky tackles that may result in serious head injury.

Table 1: Data distribution in the data set

Class Name	No. of Samples	Avg No. of Frames
Safe	123	216
Risky	55	203
Total	178	212

3 Data Set Preparation

Lack of a data set relevant to our task motivates us to construct a new data set. We attempt to solve the problem from a supervised learning point of view; therefore, we need labeled training data. We build our data set in two steps. First, we collect videos from practice fields, and then we label each video manually.

3.1 Video Collection

Our data set consists of 178 tackle videos. Originally, we collected other videos as well, but we had to discard some because of poor resolution and older encoding format. All the videos are collected from seven different practice fields in the United States. They are recorded in different formats: MOV, MOD, MKV, and MP4. All files are then converted to MP4. The frame rate for all videos is 30. A standard guideline was used to set up cameras, however, the guideline was not strictly maintained. In all videos, the player starts running from the left, and the dummy is placed on the right. Some unnecessarily long videos are trimmed to some extent.

3.2 Data Annotation

We consider the task of risky tackle identification as a binary classification problem. Therefore, we annotate each video as either 'safe' or 'risky'. The annotation is done by a certified athletic trainer who first rates every tackle on a scale of 3. Tackles scored 0 or 1 are considered risky while tackles scored 2 and 3 are considered safe. The annotator judged every video based on the head position, body posture, and contact point around the strike zone: where the player hits the dummy. Although many factors from consecutive frames are involved, loosely speaking, if the head or helmet of the player initiates the contact, the tackle is risky, but if the player uses his chest or shoulder for initial contact keeping his head away, that is considered a safe tackle. Table 1 shows the data distribution that we have after the preprocessing and annotation.

4 Approach

The main motivation behind our approach is to first extract the spatiotemporal regions that are more relevant to the task with minimum effort and then use

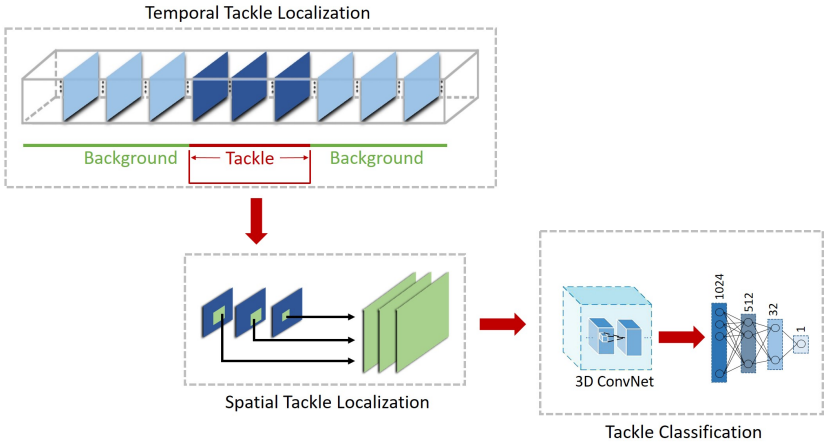

Fig. 2: Overall pipeline. In the first stage of the pipeline, tackle related frames are extracted. The second stage localizes the tackle spatially using a pre-trained Mask-RCNN model. In the final stage, spatiotemporally segmented frames pass through a 3D convolutional network and subsequent fully connected layers.

these informative segments to identify risky tackles. Figure 2 depicts the overall pipeline of our proposed approach.

4.1 Temporal Tackle Localization

Manual investigation reveals that only a few frames around the strike zone contain key information rather than frames that are more distant from the actual tackle event. More specifically, it turns out that only 10-20 frames are important where the tackle is happening compared to the huge number of frames in each video. The task of extracting action-related frames, in other words, the task of removing redundant frames, is similar to action localization. However, the drawback of considering the problem as action localization is that we need the temporal anchors or frame-level annotations defining the start and end of the tackle action in the video. Therefore, general action localization approaches require much effort for annotation. Moreover, such approaches are often multi-stage, which in turn will increase the complexity of our task.

We propose to cast the temporal tackle localization task as anomaly localization where we consider the tackle or collision with the dummy (both safe and risky) as an anomalous event. To avoid the necessity of temporal or frame-level annotations, we leverage a state-of-the-art approach [37] that uses only video-level annotation. They consider each video as a bag and video segments as instances in a deep multiple instance learning framework. To utilize such an approach, we create an auxiliary data set. From one video of the original data set, we create two videos, one before the tackle occurs and another after the

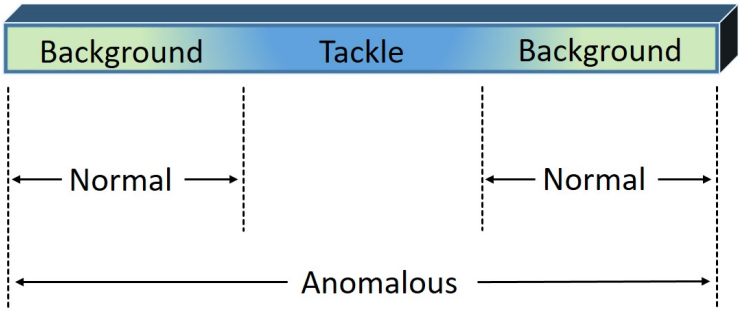

Fig. 3: Visualization of auxiliary data set creation. Our approach does not consider the highly specific start or end point of the tackle event. Any point in the gradient space avoiding the solid blue zone can be considered as the temporal anchor for the normal video.

tackle is finished, unless the tackle event is at the very start or end of the video. However, this has been done without considering the precise temporal location of the tackle, as shown in Figure 3. It just requires that the normal videos will not contain the core frames of the tackle event; thus, anyone with no domain knowledge can perform the task of video clipping. We consider these newly created videos as normal videos because they do not contain any tackle where the player is hitting the dummy. On the other hand, we use the original untrimmed videos as anomalous videos because they contain either safe or risky tackle events at some point in the video. The rationale behind using the untrimmed videos as anomalous videos instead of video of the clipped tackle event is twofold. First, the frames that do not contain the tackle event reside in both normal and anomalous videos. As [37] uses a ranking loss function that discriminates the highest scored instances in the normal and anomalous videos, the presence of non-tackle segments in both normal and anomalous videos inherently increase the score of the tackle segment. Second, at the test time, we expect our approach to identify the tackle related frames from the untrimmed videos. Thus, training the model using the whole videos resonate better with the end goal.

We train the anomaly detection model with these normal and anomalous videos, so the model learns to predict anomaly scores for each segment of a video. We propose an anomaly score-based selection mechanism for frame extraction. We run a temporal sliding window of 16 frames with 8 frame overlap and predict the anomaly score for each window. First, we select the window i with the highest anomaly score. Then the window which scores higher between the two window $i+1$ and $i-1$ is selected. Finally, we take an extra four frames before and after the two selected consecutive windows. In this way, the 32-frame long window or segment will contain the anomaly: the tackle event.

4.2 Spatial Tackle Localization

The tackle event takes place only in a particular spatial region of a frame within a large background. Figure 1 clearly shows a background containing unnecessary information such as other players, coaches, playground infrastructure, and service cart. Thus, considering only the spatial region of interest can drastically reduce the spatial dimension without the loss of any key information.

We propose to take the advantage of a pre-trained object recognition model to spatially localize the tackle event. As the tackle is performed by a person, we use the Mask R-CNN [12] object detection model to generate the bounding boxes for all persons present in the frame. The player appears in a wide bounding box because of the action performed and the camera set up in close proximity. Thus, we exploit the relative width of the bounding boxes to select the player performing the tackle when several persons are present within a frame. Finally, the selected bounding box is extended to the top and right sides to include the dummy.

4.3 Tackle Classification

We utilize a 3D convolutional network to learn the spatiotemporal features from the video frames we retain after discarding the redundant frames. Specifically, our architecture includes four 3D convolution layers each followed by a 3D max-pooling layer. The number of filters for the four convolution layers are 16, 64, 256, and 1024, respectively. We use $3\times3\times3$ convolution filters with stride $1\times1\times1$ for all layers. According to [40], to preserve temporal features in the first stage, we have a kernel size of $1\times2\times2$ and stride $1\times2\times2$ for the first pooling layer. All other 3D pooling layers are $2\times2\times2$ with stride $2\times2\times2$. A global average pooling layer connects the 3-layer fully connected (FC) block to the convolutional block. The first FC layer has 512 units, the second layer has 32 units, and the final layer has only 1 unit. ReLU activation is used for all the layers except the final one, which has Sigmoid activation. We apply 50% dropout regularization after each FC layer and use Adam optimizer with a learning rate of 0.0001. We perform parameter sweep to empirically select the best set of hyperparameters for the network. The model is trained to minimize the binary cross-entropy loss:

$$\mathcal{L} = \sum_i (y_i \; log \, p_i + (1 - y_i) \; log \, (1 - p_i)), \tag{1}$$

where y_i and p_i denotes the label and the prediction, respectively, for sample i.

5 Experimental Setup

5.1 Train-Test Split

Experiments were repeated three times with random splits of the data. In each trial, we perform a 80%-20% train-test split over the samples in the data set. In all splits, we maintain approximately the same distribution of classes as in the original data set. To ensure a robust generalizable analysis, we hold out the test set at all stages of the pipeline.

5.2 Baseline

We compare our method with one state-of-the-art video classification and one anomaly detection approach to evaluate the effectiveness of our proposed pipeline.

C3D Baseline [40]: We follow the same convention as mentioned in [40] to extract the C3D descriptor from the whole video. We obtain the fully connected (FC) layer FC6 activations of the C3D network for each 16 frame clip with an eight frame overlap. Finally, to get the C3D video descriptor, we average these clip level activations and then apply l_2 normalization that results in a 4096-dim vector. At first, we attempt to learn a Support Vector Machine (SVM) classifier using the C3D video descriptor as originally used in [40]. However, such shallow models fail to learn anything meaningful and always tend to predict the majority class. This may be due to the class imbalance and lack of sufficient representative samples from the minority class. The use of class weightage does not seem to help. Thus, we use a Multi Layer Perceptron (MLP) similar to the one described earlier in Section 4.3 as the classifier.

Anomaly Detection (AD) Baseline [37]: We compare our model with this state-of-the-art approach because the task of risky tackle detection can be considered as an anomaly detection task. We train their network assuming the risky tackles as anomalous events and the safe tackles as normal events. We use exactly the same settings for the hyperparameters as in the original implementation.

5.3 Evaluation Metrics

We consider the set of risky tackles as the positive class while safe tackles are the negative class. In the presence of class imbalance, which can be extreme for practice tackles that are supervised by coaching staff, accuracy cannot serve as an adequate figure of merit, because it does not consider the skewness in class distribution. Therefore, we report balanced accuracy, which is defined as

$$\frac{True\ Positive\ Rate + True\ Negative\ Rate}{2}.$$

Following earlier works on learning with imbalanced data sets, we also report precision, recall, and F1-score to compare the performance of risky tackle identification at the final stage. Further, we qualitatively evaluate the performance of the intermediate stages.

6 Results and Discussions

Comparison with the Baseline: Table 2 shows the quantitative experimental results for both the baselines and our proposed approach. Under the name Temporal Localization (TL) only, we report the classification results using the frames obtained just after the first stage of the pipeline. This also serve as an ablation study for the temporal localization stage. When the Spatial Localization (SL) is added on top of temporal localization, we use TL + SL to refer to it.

Table 2: Evaluation metrics for different approaches (TL: Temporal Localization, SP: Spatial Localization)

Methods	Bal. Acc.	Precision	Recall	F1-Score
C3D (untrimmed)	54.81	41.67	22.22	28.05
AD (untrimmed)	53.53	35.71	41.67	38.46
Ours (TL Only)	55.13	47.62	33.33	37.61
Ours (TL + SL)	**66.88**	**54.25**	**55.56**	**54.29**

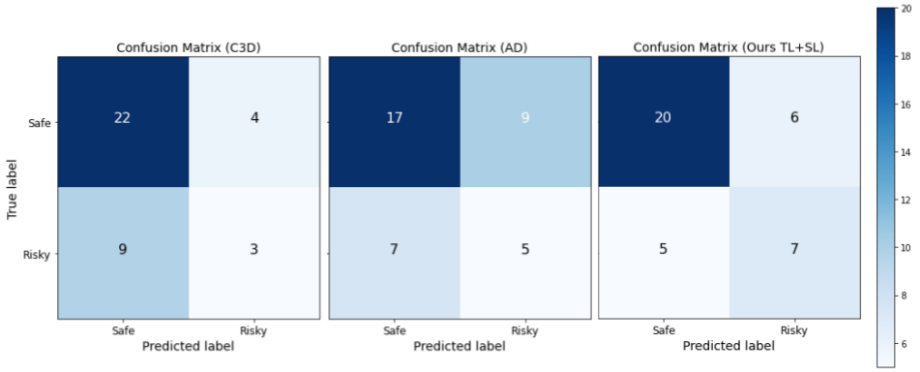

Fig. 4: Confusion matrices of a particular trial for C3D baseline (left), state-of-the-art anomaly detection model (middle), and our proposed approach (right).

Table 2 shows that our proposed 3-stage pipeline outperforms the C3D and Anomaly Detection (AD) based approaches applied to the untrimmed videos in terms of all metrics. Our approach achieves $12 - 13\%$ higher balanced accuracy than the other approaches, which means it detects both classes better. In particular, our 2-stage (TL only) approach achieves 6% higher precision than the C3D baseline and and 12% higher precision than the AD baseline. It further improves by 6% when combined with spatial localization. In terms of recall, our TL only approach does not do better than the AD baseline, however, the 3-stage (TL+SL) approach achieves significantly higher recall than C3D and outperforms the AD baseline by a considerable margin. Moreover, our (TL+SL) approach is able to achieve an F1-score of 54.29, which is 26% and 15% better than the C3D and AD baselines, respectively. Therefore, our (TL+SL) approach achieves higher recall and F1-score compared to the other approaches without compromising precision. This denotes the effectiveness of our approach in detecting the positive risky tackle class and maintaining a good balance between positive and negative class detection. Figure 4 presents the confusion matrices for the test set of a particular trial. Our model shows similar time complexity compared to the AD baseline, however, slightly higher than the C3D baseline. The main computational bottleneck stems from the temporal localization stage.

Table 3: Ablation study for temporal and spatial localization

Methods	Precision	Recall	F1-Score
C3D (untrimmed)	41.67	22.22	28.05
C3D (TL Only)	42.86	25.00	30.90
C3D (TL + SL)	52.22	16.67	25.07
Ours (MT Only)	**60.32**	27.22	36.96
Ours (MT + SL)	49.02	30.56	31.49
Ours (TL Only)	47.62	33.33	37.61
Ours (TL + SL)	54.25	**55.56**	**54.29**

Ablation Study: We perform an ablation study to analyze the necessity and efficacy of the intermediate stages. We also manually localize the tackle for robust comparison and report the results using the notation Manually Trimmed (MT). We answer the following research questions to interpret the experimental results:

– *Is there any improvement in performance due to the temporal localization?*

As we can see from Table 3, the TL and MT only approaches always outperform the C3D baseline for untrimmed video in all aspects. We have extracted 32 frames while the average number of frames in untrimmed videos is 212. Thus, removing a significant portion of the frames does not affect the performance negatively rather improves it and saves computation power. Also, this reduction in the number of frames opens up the possibility of learning spatiotemporal features directly from all frames instead of averaging the features obtained from chunked video clips.

– *Does the use of the spatial localization improve the classification performance?*

Table 3 shows that the addition of the spatial localization stage always increases the recall for our proposed model compared to the TL or MT only counterparts. That means removing unnecessary spatial information contributes largely to improve the detection of risky tackles. However, the C3D model performs poorly when combined with spatial localization, possibly because of the biases of the pre-trained C3D features towards unsegmented frames.

Accuracy of Temporal Localization: We manually evaluate whether tackles are present in the 32-frame long clips extracted from all the test videos. It turns out that the localization model trained in stage one achieves 97% accuracy in the test set for the temporal localization task. Therefore, we conclude subsequent stages will require more attention to improve the overall classification performance.

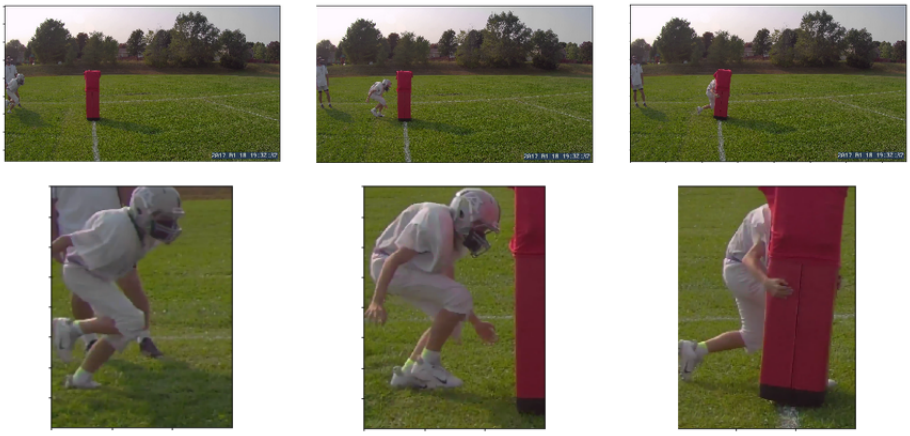

Fig. 5: Example of spatial localization. The top row shows the original frames and the bottom row presents the corresponding spatially segmented ones.

Accuracy of Spatial Localization: We leveraged the Mask-RCNN model from the Detectron2 [43] library in such a way that the detection of the player is guaranteed in most cases. Figure 5 presents some examples of spatial localization. However, when the player has not entered the camera's field of view and there is another person in the frame, our method occasionally selects that person as a player. Also, just after the tackle when the player trends downward with the dummy, the player is often occluded by the dummy. This may cause the spatial localization model to fail. However, such failures may not affect the detection task significantly, because these scenarios arise either before the tackle event has started or after hitting the dummy.

7 Conclusions

In this paper, we propose a 3-stage pipeline to detect risky tackles from American football practice videos. The experimental results show that our proposed method performs significantly better than the existing 3D ConvNet-based methods for video classification. There are limitations due to the size of our presented data set, however, the inherent skewness poses a substantial challenge for improving the model performance even beyond the random guess. In the future, we would like to use multi-modal approaches to take the benefit of optical flow features and pose estimation. The use of this automatic risky tackle identification framework can provide faster feedback to the player, and such feedback and supervision during the tackle practice can significantly minimize the risks of head impact and head injuries among young American football players.

Acknowledgements. We would like to thank Ademola Okerinde for his insightful discussions and Nazmun Akter Pia for her assistance in creating the visualizations.

References

1. Alosco, M., Kasimis, A., Stamm, J., Chua, A., Baugh, C., Daneshvar, D., Robbins, C., Mariani, M., Hayden, J., Conneely, S., et al.: Age of first exposure to american football and long-term neuropsychiatric and cognitive outcomes. Translational psychiatry **7**(9), e1236–e1236 (2017)
2. Buch, S., Escorcia, V., Shen, C., Ghanem, B., Carlos Niebles, J.: Sst: Single-stream temporal action proposals. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pp. 2911–2920 (2017)
3. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6299–6308 (2017)
4. Chrisman, S.P., Lowry, S., Herring, S.A., Kroshus, E., Hoopes, T.R., Higgins, S.K., Rivara, F.P.: Concussion incidence, duration, and return to school and sport in 5- to 14-year-old american football athletes. The Journal of pediatrics **207**, 176–184 (2019)
5. Cournoyer, J., Tripp, B.L.: Concussion knowledge in high school football players. Journal of athletic training **49**(5), 654–658 (2014)
6. Cui, X., Liu, Q., Gao, M., Metaxas, D.N.: Abnormal detection using interaction energy potentials. In: CVPR 2011, pp. 3161–3167. IEEE (2011)
7. Dompier, T.P., Kerr, Z.Y., Marshall, S.W., Hainline, B., Snook, E.M., Hayden, R., Simon, J.E.: Incidence of concussion during practice and games in youth, high school, and collegiate american football players. JAMA pediatrics **169**(7), 659–665 (2015)
8. Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., Darrell, T.: Long-term recurrent convolutional networks for visual recognition and description. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2625–2634 (2015)
9. Escorcia, V., Heilbron, F.C., Niebles, J.C., Ghanem, B.: Daps: Deep action proposals for action understanding. In: European Conference on Computer Vision, pp. 768–784. Springer (2016)
10. Giancola, S., Amine, M., Dghaily, T., Ghanem, B.: Soccernet: A scalable dataset for action spotting in soccer videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 1711–1721 (2018)
11. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 580–587 (2014)
12. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision, pp. 2961–2969 (2017)
13. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778 (2016)
14. Heilbron, F.C., Niebles, J.C., Ghanem, B.: Fast temporal activity proposals for efficient detection of human actions in untrimmed videos. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1914–1923 (2016)

15. Hospedales, T., Gong, S., Xiang, T.: A markov clustering topic model for mining behaviour in video. In: 2009 IEEE 12th International Conference on Computer Vision, pp. 1165–1172. IEEE (2009)
16. Ji, S., Xu, W., Yang, M., Yu, K.: 3d convolutional neural networks for human action recognition. IEEE transactions on pattern analysis and machine intelligence **35**(1), 221–231 (2012)
17. Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Fei-Fei, L.: Large-scale video classification with convolutional neural networks. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pp. 1725–1732 (2014)
18. Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., et al.: The kinetics human action video dataset. arXiv preprint arXiv:1705.06950 (2017)
19. Kelley, M., Urban, J., Jones, D., Powers, A., Whitlow, C.T., Maldjian, J., Stitzel, J.: Football concussion case series using biomechanical and video analysis. Neurology **91**(23 Supplement 1), S2–S2 (2018)
20. Kelley, M.E., Jones, D.A., Espeland, M.A., Rosenberg, M.L., Miles, C.M., Whitlow, C.T., Maldjian, J.A., Stitzel, J.D., Urban, J.E.: Physical performance measures correlate with head impact exposure in youth football. Medicine and science in sports and exercise **52**(2), 449 (2020)
21. Kelley, M.E., Kane, J.M., Espeland, M.A., Miller, L.E., Powers, A.K., Stitzel, J.D., Urban, J.E.: Head impact exposure measured in a single youth football team during practice drills. Journal of Neurosurgery: Pediatrics **20**(5), 489–497 (2017)
22. Kerr, Z.Y., Roos, K.G., Djoko, A., Dalton, S.L., Broglio, S.P., Marshall, S.W., Dompier, T.P.: Epidemiologic measures for quantifying the incidence of concussion in national collegiate athletic association sports. Journal of athletic training **52**(3), 167–174 (2017)
23. Kerr, Z.Y., Yeargin, S., Valovich McLeod, T.C., Nittoli, V.C., Mensch, J., Dodge, T., Hayden, R., Dompier, T.P.: Comprehensive coach education and practice contact restriction guidelines result in lower injury rates in youth american football. Orthopaedic journal of sports medicine **3**(7), 2325967115594,578 (2015)
24. Klaser, A., Marszałek, M., Schmid, C.: A spatio-temporal descriptor based on 3d-gradients. In: BMVC 2008-19th British Machine Vision Conference, pp. 275–1. British Machine Vision Association (2008)
25. Lai, K.T., Liu, D., Chen, M.S., Chang, S.F.: Recognizing complex events in videos by learning key static-dynamic evidences. In: European Conference on Computer Vision, pp. 675–688. Springer (2014)
26. Lai, K.T., Yu, F.X., Chen, M.S., Chang, S.F.: Video event detection by inferring temporal instance labels. In: Proceedings of the ieee conference on computer vision and pattern recognition, pp. 2243–2250 (2014)
27. Laptev, I.: On space-time interest points. International journal of computer vision **64**(2-3), 107–123 (2005)
28. Lessley, D.J., Kent, R.W., Funk, J.R., Sherwood, C.P., Cormier, J.M., Crandall, J.R., Arbogast, K.B., Myers, B.S.: Video analysis of reported concussion events in the national football league during the 2015-2016 and 2016-2017 seasons. The American journal of sports medicine **46**(14), 3502–3510 (2018)
29. McAllister, T., McCrea, M.: Long-term cognitive and neuropsychiatric consequences of repetitive concussion and head-impact exposure. Journal of athletic training **52**(3), 309–317 (2017)

30. Oneata, D., Verbeek, J., Schmid, C.: Action and event recognition with fisher vectors on a compact feature set. In: Proceedings of the IEEE international conference on computer vision, pp. 1817–1824 (2013)
31. Piergiovanni, A., Ryoo, M.S.: Early detection of injuries in mlb pitchers from video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 0–0 (2019)
32. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. arXiv preprint arXiv:1506.01497 (2015)
33. Scovanner, P., Ali, S., Shah, M.: A 3-dimensional sift descriptor and its application to action recognition. In: Proceedings of the 15th ACM international conference on Multimedia, pp. 357–360 (2007)
34. Shou, Z., Wang, D., Chang, S.F.: Temporal action localization in untrimmed videos via multi-stage cnns. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1049–1058 (2016)
35. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. arXiv preprint arXiv:1406.2199 (2014)
36. Smart, B.J., Haring, R.S., Asemota, A.O., Scott, J.W., Canner, J.K., Nejim, B.J., George, B.P., Alsulaim, H., Kirsch, T.D., Schneider, E.B.: Tackling causes and costs of ed presentation for american football injuries: a population-level study. The American journal of emergency medicine **34**(7), 1198–1204 (2016)
37. Sultani, W., Chen, C., Shah, M.: Real-world anomaly detection in surveillance videos. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 6479–6488 (2018)
38. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2818–2826 (2016)
39. Tierney, G.J., Kuo, C., Wu, L., Weaving, D., Camarillo, D.: Analysis of head acceleration events in collegiate-level american football: A combination of qualitative video analysis and in-vivo head kinematic measurement. Journal of Biomechanics **110**, 109,969 (2020)
40. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3d convolutional networks. In: Proceedings of the IEEE international conference on computer vision, pp. 4489–4497 (2015)
41. Wang, H., Schmid, C.: Action recognition with improved trajectories. In: Proceedings of the IEEE international conference on computer vision, pp. 3551–3558 (2013)
42. Wu, L.C., Kuo, C., Loza, J., Kurt, M., Laksari, K., Yanez, L.Z., Senif, D., Anderson, S.C., Miller, L.E., Urban, J.E., et al.: Detection of american football head impacts using biomechanical features and support vector machine classification. Scientific reports **8**(1), 1–14 (2017)
43. Wu, Y., Kirillov, A., Massa, F., Lo, W.Y., Girshick, R.: Detectron2. https://github.com/facebookresearch/detectron2 (2019)
44. Xu, D., Ricci, E., Yan, Y., Song, J., Sebe, N.: Learning deep representations of appearance and motion for anomalous event detection. arXiv preprint arXiv:1510.01553 (2015)
45. Zhao, B., Fei-Fei, L., Xing, E.P.: Online detection of unusual events in videos via dynamic sparse coding. In: CVPR 2011, pp. 3313–3320. IEEE (2011)