Highlight Extraction from Football Match Video

Minseo Kim
Department of Computer Science
Soongsil University
superider00@gmail.com

1 Introduction

Manually creating highlight reels from sports matches is a time-consuming process. This project aims to automate the generation of highlights in football by detecting and extracting important game events using computer vision and deep learning. By identifying key events like goals or fouls and associating them with time segments, this project can create summary videos with minimal human intervention.

This Project proposes a deep learning-based computer vision system for automatically generating highlights from full-length football match videos. The system is designed to detect key match events such as goals, penalty kicks, and corner kicks, or shots on target and shots off target. My plan is to train the RNN-based Model with extracted features from retrained ResNet50 on the Soccernet dataset. The trained model will be learn which frame should be extracted from a input video. After extracting the short cuts, the output video will be well-made highlight video.

2 Summary of Questions and Results

1. **Can it perform better than currently published video highlight extraction models?**

Expected Result : Since most of the current video highlight extraction model is not a model trained only in a specific domain, the performance may not be good compared to the current project model that is scheduled to be trained only with training data in that domain.

Determined Result : To get the model to extract the highlights of the football game as I intended, training with the data only with the football domain was needed and it actually focused on the content of the football game.

2. **Can deep learning models accurately detect key football events such as goals, fouls, and corner kicks from broadcast footage?**

Expected Result : This project will investigate the video flow understanding performance of RNN-based models trained on labeled football events.

Determined Result : I simplify the detecting-each-event task to the highlight and non-highlight binary classification task. Since my ultimate goal was just extracting

the highlight video from a raw video, I was able to simplify the problem like that, and it actually get great accuracy as I intended.

3. Can the detection results be effectively converted into watchable highlight clips?

Expected Result : This project is planning to verify that the video of natural results as much as possible can be effective according to the existing purpose using a properly learned model and ffmpeg.

Determined Result : This project's model is mainly trained with GRU, which has not many parameters to learn compared by the transformers which is proposed mostly in the current published video highlight extraction model.

3 Motivation

In general video understanding, especially in case of event detection or action detection, most of them use transformer or CNN based architecture for video task like 3D CNN to extract time-space feature. After that, they fusion the multi-modal data like audio data for extracting the highlight scene from input video. Also, prior works focus primarily on video understanding regardless with specific domain. Even in sports(football) field, prior works focus on player and ball tracking. Projects' approach can help extracting highlight specifically and easily from football video. I thought fine-tuning with existing video highlight extraction model is inevitable to perform better in the soccer domain than the existing video highlight extraction model. However, these models are using complex models and needs much additional time. With the aim of creating the most efficient and accurate model within the single domain of football, judging that it would be too inefficient to use the current models, I propose my model to extract the highlight video from the full football video accurately and both efficiently.

4 Dataset

Soccernet dataset (https://www.soccer-net.org/) will serve as the primary source of annotated football events. Additional data didn't use because of the lack of storage space. Each videos are annotated for classes: goal, foul, corner kick, penalty kick. It includes over 500 full-match videos with labeled events (goals, fouls, penalties, etc.). Videos available upon NDA approval. Event timestamps and metadata are provided as JSON files.

It contains roughly 50 ~ 100 full videos each league, and each game divided into the 4 videos, first-half and the second-half, 224p and 720p. Since I didn't own any GPU for myself and much amount of storage needs to contain those datasets, I used Google Drive and Google Colab to train my model and inference it.

Train (80%) / Validation (20%) : First-half video (224p) for total 480 games (England 95 games, France 40 games, Germany 53 games, Italy 96 games, Spain 95 games, Europe (Uefa Champions League) 101 games)

Test : One second-half video per each inference (because of the lack of storage space)

5 Method

Dataset Configuration and Training Data Generation (Preprocessing)

Using almost 500 first and second half videos, I needed about 200GB or more, so I used only the first half of 224p (25fps) to reduce weight. I tried manual labeling at first, but as the Soccernet offer the labeled file, I just added the highlight and non-highlight binary label on the file and extract the divided 15-second clips which is split into highlight and non-highlight classes. I regarded the 'Goal', 'Penalty', 'Shots off target', 'Shots on target' annotated clips as a highlight clips.

Feature extraction and Highlight Classification Training

Before training, I augmented the video datas using random resize, random rotation, gaussian blur and normalize them.

For the training, Yolov5 and Faster R-CNN-based object detection method considered at the beginning was judged to be inappropriate. Since the core of Action Detection is a time flow and one of the project's goal was to make the model efficient, I embed each clip's frame through CNN (ResNet50) for feature extraction on a frame level and then put those clips in the RNN-based model(GRU) in chronological order to make my model learn what kinds of clips are highlight or non-highlight.

Highlight Detection Evaluation and Post Processing

Highlight output video creation and frame extraction conducted based on the results classified by the model. For a variety of uses, each extracted clips of highlights in addition to full highlight video are the model output.

6 Results

1) Research results

1. **Can it perform better than currently published video highlight extraction models?**

When I uploaded the full football video over the current published video highlight extraction model, it focused on the close-ups of individual players or coaches, and scenes that illuminated the audience, not from the perspective of football game content. To get the model to extract the highlights of the football game as I intended, training with the data only with the football domain was needed and it actually focused only on the content of the football game.

2. **Can deep learning models accurately detect key football events such as goals, fouls, and corner kicks from broadcast footage?**

As I design my own model architecture and training my model to learn the football events such as goals, fouls, and corner kicks, it was expected to be able to get a great accuracy of detecting out the scenes exactly each goal, foul, shoots on target and corner kick. But as I think of the multi-classification task, my model should

been more heavy to get a great accuracy. So instead, I simplify the detecting-each-event task to the highlight and non-highlight binary classification task. Since my ultimate goal was just extracting the highlight video from a raw video, I was able to simplify the problem like that, and it actually get great accuracy as I intended.

However, it's not yet clear whether the model is well built so it cause the expected result. Due to the nature of the football video, many similar scenes appear near both goal lines, so I thought that it might have been learned without difficulty regardless of model construction.

3. Can the detection results be effectively converted into watchable highlight clips?

I extract only 75 frames(only 5 frame each second) for training instead of the full 15 second video input I preprocessed. Also, this project's model is mainly trained with GRU, which has not many parameters to learn compared by the transformers which is proposed mostly in the current published video highlight extraction model. This contributes this project to assert about the efficiency of the model within the football video highlight extraction field.

In addition, as a kind of short form for one highlight scene has also become popular these days, allowing the output to come out in the form of a short clip seem to be done as well as a full highlight video.

2) Result Interpretation

As a result of loss, training accuracy and validation accuracy for each epoch during training, it was found that the overall loss decreased gradually and the accuracy increased. A total of 40 epochs were performed, and the training accuracy was steadily increasing, whereas the validation accuracy was fell after 27 epochs. It seems to be overfitted after 25 ~ 30 epoch. Since I stored weights for each epoch, I was able to conduct an inference with the weights at the time of epoch 27, which showed the best validation performance.

[Loss & Accuracy of each training epoch]

```
Epoch [1/40]
Train Loss: 0.6391, Train Accuracy: 62.78%
Valid Loss: 0.6061, Valid Accuracy: 69.03%
```

```
Epoch [6/40]
Train Loss: 0.5788, Train Accuracy: 71.81%
Valid Loss: 0.5427, Valid Accuracy: 76.11%
```

```
Epoch [11/40]
Train Loss: 0.5317, Train Accuracy: 77.31%
Valid Loss: 0.5013, Valid Accuracy: 79.65%
```

```
Epoch [16/40]
Train Loss: 0.4863, Train Accuracy: 82.38%
Valid Loss: 0.4953, Valid Accuracy: 80.53%
```

```
Epoch [21/40]
Train Loss: 0.4250, Train Accuracy: 88.77%
Valid Loss: 0.4466, Valid Accuracy: 84.96%
```

```
Epoch [27/40]
Train Loss: 0.3968, Train Accuracy: 91.63%
Valid Loss: 0.4099, Valid Accuracy: 90.27%
```

```
Epoch [40/40]
Train Loss: 0.3595, Train Accuracy: 95.37%
Valid Loss: 0.4567, Valid Accuracy: 84.96%
```

7 Impact

Since the football is a sport that is wild all over the world, fans who follow many matches have to watch those matches early morning hours. It's often difficult to watch the game at dawn. This project can help those people watch the favorite match without much waste of time during the day. This project can also give a big reduction of time for spending manually editing highlight reels. Fans and viewers usually wait for the highlight video to be uploaded as quick as possible. And so, editorial staff is not given much time. Because the full football match is about 2 hours long, someone who works for the sports technique company or sports broadcast/editing system can reduce their working hours and make the viewers satisfied by this project's model.

But in the case of my current model, since labels such as goal, shots on target, and shots off target are set as highlights, it may be misunderstood that a scene which is not in the output of the model is an insignificant scene. Also, I don't think it can be such a good model for fans who want to see more elements outside the match such as the team's coach and bench or other events during the match.

8 Challenge Goals

This project was aimed to meet the goal to create the **most efficient, accurate highlight extraction model** that can work **in the football domain**. This project is designed to result in a fully functional system that can process full-length football match videos and automatically generate coherent highlight clips. By combining pre-trained CNN (resnet50) feature and RNN based model with the automated video editing tools (such as ffmpeg), this project makes us to expect to deliver a tangible and practical outcome (a highlight video that can be viewed and evaluated directly). I think the goal to create the **most efficient, accurate highlight extraction model** that can work **in the football domain** is justified because RNN is a model that is not heavier than transformer-based models, and due to the specificity of the domain of football, similar scenes flow is often. Therefore, I judged that RNN alone could perform well enough without modules such as attention mechanisms. As a result, meaningful results has been expected even if only the frame and time level are given. And so it can be considered efficient.

From a big perspective, my first challenge goal has not changed, and it seems to have been completed well as I intended.

9 Work Plan Evaluation

Expected Plan : Training will be conducted on a dataset of at least 500 videos. Dataset collection and manual annotation(labeling) will be cost a long time. If it is determined that a scene that is too absurd is included after training and evaluating, I am planning to increase the data or modify the model structure.

Task 1 - Dataset collection and manual annotation - 12h

Task 2 - Model training - 30h (few days)

Task 3 - Model evaluation and benchmarking - 12h

Task 4 - Frame extraction and metadata generation - 8h

Task 5 - Highlight video creation using ffmpeg - 10h

Executed Work : As I expected, dataset collection and labeling has cost a long time. But in addition, implementing the data loader for this project which was not considered to be a difficult cost a lot of time. Since it's my first time seeing a video task and I'm not familiar with the dataset, preprocessing and loading video frames on my model was not an easy process. Before training, aligning the tensor dimensions of the data was also a troublesome task for me. Therefore, Task 1 described before in Expected Plan has taken much a lot more time than I expected (about 2 days). Because the Task 4 and 5 was just for a inference, I would remove from the work execution. Actually, training and model evaluation after training has taken amount of time exactly I expected. I think the lightweight of the model helps me to estimate close to reality, and my lack of experience and coding skills makes my estimates far from reality.

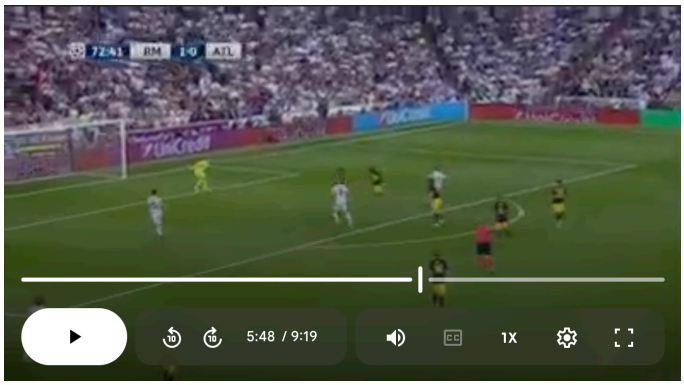
10 Testing

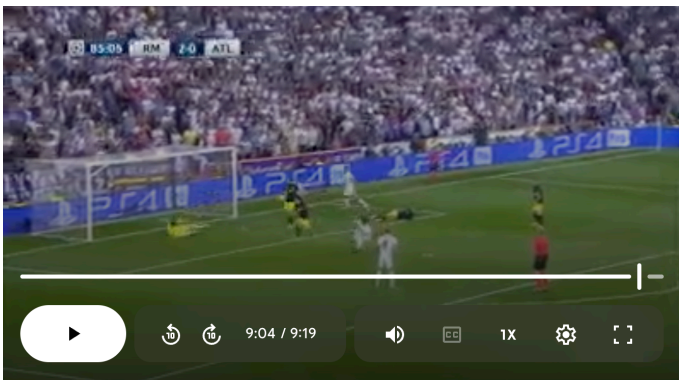
I test my model about 20 match using second-half videos. Because of the lack of storage, I had to upload each video each time I performed an inference. Testing dataset was also can be provided from Soccernet.

For a following inference report, I used two example using one of the match from Uefa Champions League named "2017-05-02 - 21-45 Real Madrid 3 - 0 Atl. Madrid" - [1], and "2015-05-17 - 18-00 Manchester United 1 - 1 Arsenal" - [2] from England Premier League. I was able to realize that every goals and fouls, and corner kicks has been concluded in the output video. However, I was also able to realize that few touch-out scenes (ball out of the side line) were considered as highlight. I regard that the model recognize that scene as a penalty or shots off target.

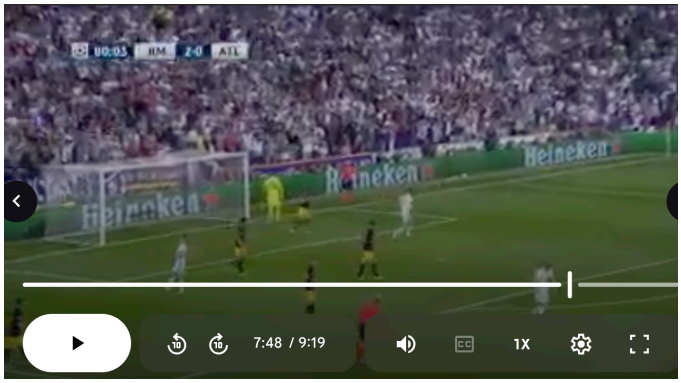
[Shots On Target] - 2 examples of [2] match

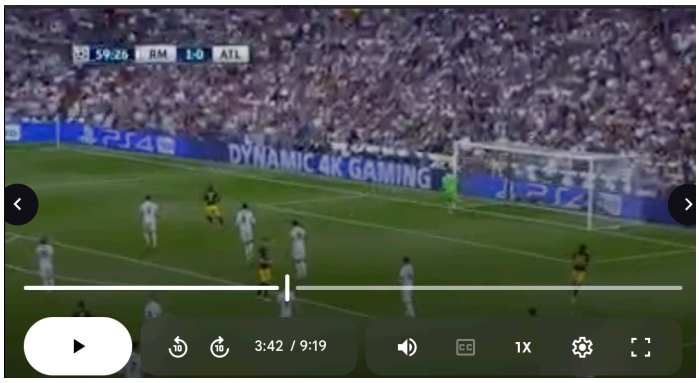
[Goals] - 2 examples of [1] match, 1 example of [2] match

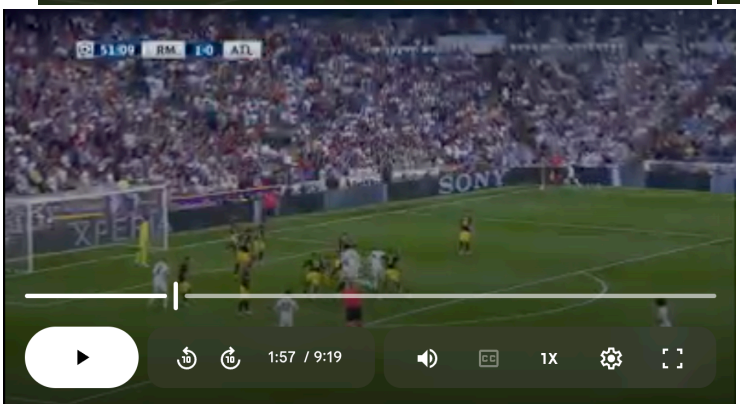




[Shots Off Target] - 3 examples of [1] match, 1 example of [2] match







As I test my model using the 224p video, output can look bit blurry. Examples above on **[Shots On Target]** were shooting scene in which the opponent's goalkeeper catches it. Examples above on **[Goals]** were shooting scene right before the goal. Examples above on **[Shots Off Target]** were the scene right after the shooting off the post.

11 Collaboration & Related Works

I used Google Drive and Google Colab to train my model.

I searched/referred the resources below.

<Train / Test dataset>

Soccernet - https://www.soccer-net.org/

<GRU model>

Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation (https://arxiv.org/abs/1406.1078)

<Proposed Related Papers Before>

Automated Detection of Sport Highlights from Audio and Video Sources (https://arxiv.org/abs/2501.16100, https://github.com/ChakradharG/AuViS)

Detecting key Soccer match events to create highlights using Computer Vision (https://arxiv.org/abs/2204.02573)

Automatic Soccer Game Highlight Detection (https://cs231n.stanford.edu/2024/papers/automatic-soccer-game-highlight-detection.pdf)

SPNet: A deep network for broadcast sports video highlight generation(https://www.sciencedirect.com/science/article/abs/pii/S0045790622000817)