Automatic Highlight Generation of Soccer Videos

1st Jiwoo Park
AI Graduate School
Pohang University of Science and Technology
Pohang, Republic of Korea
jwp0728@postech.ac.kr

1st Younkyung Jwa
AI Graduate School
Gwangju Institute of Science and Technology
Gwangju, Republic of Korea
whkdbsrud12@gm.gist.ac.kr

1st Jiin Kwak
AI Graduate School
Ulsan National Institute of Science and Technology
Ulsan, Republic of Korea
jiin1938@unist.ac.kr

1st Jihun Lim
Department of Mathematics
Korea University
Seoul, Republic of Korea
rimjason@korea.ac.kr

1st Sehee Kim
Department of Computer Science and Engineering
Seoul National University
Seoul, Republic of Korea
mihee90@snu.ac.kr

***Abstract*—Due to soccer's immense popularity, numerous soccer matches are played each year and people watch or record these matches. However, considering that the duration of a game of soccer is at least 90 minutes and that important scenes such as goals and passes take up a small portion of the total time, we inevitably look for highlights rather than the entire game. Therefore, we propose a model that automatically generates a highlight video of a full soccer match in this paper. For this purpose, we searched diverse action-spotting models to choose a backbone model for our highlight classifier. Through various experiments, NetVLAD++ is selected as a backbone model and we created a fairly natural highlight generator by combining NetVLAD++ and our highlight classifier. Our results are meaningful because they can be applied in various fields and have huge room for improvement.**

***Keywords*—Highlight Generation, Automatic Video Summarization, Soccer Video Highlight**

I. Introduction

In 2022, YouTubers uploaded about 720,000 hours of fresh video content every day, which means we need 82 years and 70 days to watch all of those videos made in just one day. Because there are a lot of other video platforms such as Netflix and Vimeo, we cannot deny we are living in a flood of videos. Due to this trend, the necessity of automatic video summarization has been gradually increasing [1]–[6].

We propose a highlight generation model focusing on soccer videos, which is the most popular sport among numerous sports videos. Each soccer game lasts longer than 90 minutes. However, the duration of the highlight video for one soccer match is less than a tenth of that. According to a Korean press article, it takes about 13 hours for a video editor to make 10 minutes of a highlight video from a 90 minutes video. We could save a lot of time if there was an automatic highlight generator. Another real-world usage is that automatic highlight generation can be utilized in analyzing the players of a tremendous number of soccer teams. Therefore, the creation of an automatic highlight generator for sports games will be very useful. From this intuition, we decided to create an Automatic Highlight Generation Model for soccer videos using the SoccerNet-v2 [7].

However, automatic highlight generation, which can be regarded as an automatic video summarization method, might be a very difficult task since it requires huge amounts of data and the performance of previous work is not enough to be used as a practical application. Even the action spotting problem that points out current actions (e.g., shooting, drawing, etc.) has not been solved with high accuracy considering the Mean Average Precision (mAP) score of the state-of-the-art action spotting model is 53.4.

Therefore, we instead propose a highlight classification model. From the action spotting model, we created a highlight classification model by defining the replayed parts of the video as a highlight. We define highlights as replayed parts for the following reasons. First, highlight videos include diverse actions such as goals, passes, and dribbling. Therefore, it is difficult to define a highlight scene just by using action information. Hence, we chose replayed parts from the full video as highlights as important scenes are always replayed. The highlight classification model (replay classification model) can be trained based on feature extraction and fine-tuning with additional layers. After that, we use the mAP score as a metric to evaluate the performance of our models. To choose the best-performing model, we implemented diverse structured models. Through these attempts, we created a well-performing automatic highlight classification model.

The most important part of a soccer game is the goal scene. Therefore, we created a highlight video by merging

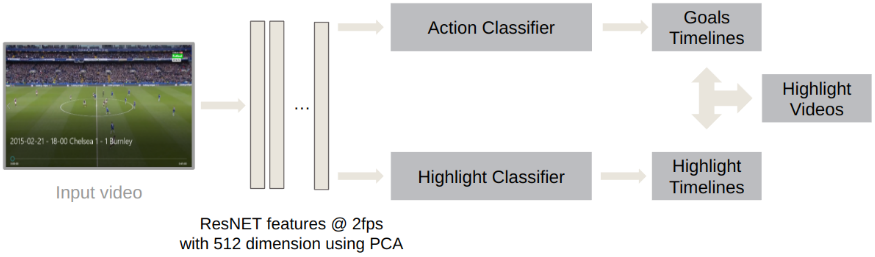


Fig. 1: Overall structure of our pipeline

goal action classification and replay highlight detection. Fig. 1 is the overall structure of our model and it will be further explained in section III.

II. Related Works

A. Video Action Detection

Video action detection is a fundamental task in computer vision that needs to recognize and localize the action performed in a video clip. It significantly impacts applications such as robotics, security, and health [8]. Region convolutional 3D(R-C3D) network proposed the end-to-end learning framework in various lengths of videos [9]. While previous activity detection approaches relied heavily on sliding window techniques [10], [11], R-C3D proposed a novel model that combines proposal and classification steps, and leverages fully connected C3D features shared between these two parts of the network, resulting in improved performance. Tran et al. [12] suggested the R(2+1)D convolution block using residual networks which combines the advantages of 2D and 3D convolutions. More recently, transformer based TimeSformer is proposed to consider both spatial and temporal features in the video [13]. Also, Slowfast introduced the architecture including two pathways: (1) a slow pathway, which is applied to low-frequency rate frames to capture the spatial semantics, and (2) a fast pathway, which is implemented to high-frequency rate frames to capture the details of the video [14].

B. Action Spotting

Giancola et al. [15] proposed action spotting as a task of finding the anchor time of a specific event in a video, rather than classifying or localizing actions. They organized the SoccerNet Action Spotting Challenge in 2021 and 2022, which attracted various models. The NetVLAD model is a CNN architecture that incorporates a differentiable pooling layer with DNN and is known for its ease of application to other CNN structures, with the differentiable pooling layer placed before the classification layer [16]. Context-Aware Loss Function (CALF) proposed a new loss function that captures the temporal context well [17]. NetVLAD++ used pooling methods considering the temporal context as a single set to pool from unlike previous papers using pooling methods, it divides context into two parts: previous to the action and succeeding to the action [18]. This makes it a better-performing model than previous work.

In the SoccerNet Action Spotting Challenge 2022, Yahoo Research got the first rank [19]. Most of the participants in the challenge proposed encoders based on transformers and learned spatial and temporal self-attention mechanisms. Yahoo Research's anchor-based method achieved the highest performance in terms of loose and tight average-mAP, but PTS outperformed in the subset of visible actions, leveraging its focus on visual cues. Yahoo Research's method and PTS outperformed the baseline model provided by the SoccerNet Challenge 2021, achieving higher tight average-mAP scores of 67.81 and 66.73, respectively, compared to the baseline's 49.56. Both approaches also demonstrated higher loose average-mAP performance.

C. Backbone Selection

Our goal is to extract specific highlights from long soccer videos. An action spotting model is used as a backbone to form the highlight classification model. Highlight classification is performed by applying two methods, feature extraction, and fine-tuning, using a pre-trained action spotting model. Our highlight extraction pipeline also makes use of goal action classification. Therefore, we conducted prior research to find a suitable action spotting model.

We chose CALF and NetVLAD++ as baseline models for the action spotting using SoccerNet-v2 data. The Average-mAP metric was utilized to compare the performance of each model. According to previous studies, the average-mAP of the CALF model was 40.7, and the average-mAP of NetVLAD++ was 53.4 [18]. In particular, NetVLAD++ recorded a higher average-mAP score than CALF for all labels except the Goal label. We tested the code introduced in the paper to see if this result appeared.

In Table I, we compare our implementation of CALF and NetVLAD++ and present the average-mAP values for each action class and the average scores for all classes. The model performs multi-label classification for each of the 17 action classes, including goal. We extracted SoccerNet-v2 data features using ResNet-152 at 2fps. NetVLAD++ outperformed CALF in terms of performance, and the precision for red and yellow card actions was very low, likely due to their infrequent occurrence in the training videos. The CALF model achieved an average-mAP of 38.9, while NetVLAD++ achieved an average-mAP of 51.0, which was similar to the original experimental results. As a result, we chose NetVLAD++ as the baseline model.

III. Materials and Methods

A. Pipeline

The pipeline can be described as follows: First, we extract features from the input video. Then, we use two types of classifiers to identify the timeline of highlights. The action classifier is utilized to detect the goal timeline, while the highlight classifier is used to identify the replay timeline. By combining these timelines, we obtain the final highlight video. Fig. 1 depicts the entire structure of our pipeline. In the upcoming section, we will elaborate on preprocessing, highlight classification, and post processing.

B. Preprocessing

We used the extracted features which are provided by SoccerNet-v2. The process for obtaining features is as follows. First, we extract the video frame images into 2fps with a resolution of 224 × 224. After that, we extract the features from the cropped videos of SoccerNet-v2 using ResNet-152, which is pre-trained on ImageNet [20]. The feature size after the ResNet-152 is 2048 and we apply PCA to reduce the feature size into 512. The number 512 was chosen by comparing the results with and without PCA. To apply video features to the NetVLAD++ model, we utilized the same window chunk approach as described in the paper. Specifically, we divided the continuous video features into chunks of equal length, with each chunk consisting of video frames extracted every T seconds.

For action spotting, we used a multi-label one-hot encoding to label all actions that occurred within a chunk. On the other hand, for highlight classification, a binary label is assigned to indicate whether a highlight was present in each chunk or not.

C. Highlight Classification

The highlight classification model uses NetVLAD++ as a backbone model. For this, we used two methodologies: feature extraction and fine-tuning [21]. Each video chunk is labeled according to highlight and has three labels. During training, we used a binary cross-entropy as a loss function for each class.

By using feature extraction, we can train other learning layers like multi-layer perceptron (MLP) with another label. We used the extracted features using the weights of the NetVLAD++ model which is pre-trained with action classes. The feature is extracted from NetVLAD++ pooling layers, which have 32,768 dimensions. Then, we trained the model using MLP with highlight labels and optimized the binary cross-entropy loss for each class.

The fine-tuning method uses the same model structure as feature extraction but differs in whether pre-trained weights are trained or not. We use the weights up to the pooled feature layer of the NetVLAD++ model trained by action spotting labels. Then, the dimension of the last fully connected layer is changed from 18 to 3 and the whole model structure is re-trained using highlight labels. We also use the binary cross-entropy loss for each class.

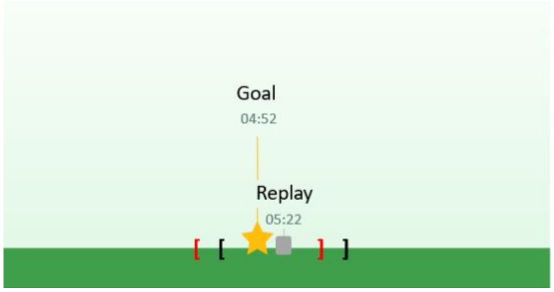


Fig. 2: Selecting the Run Time of the Highlights

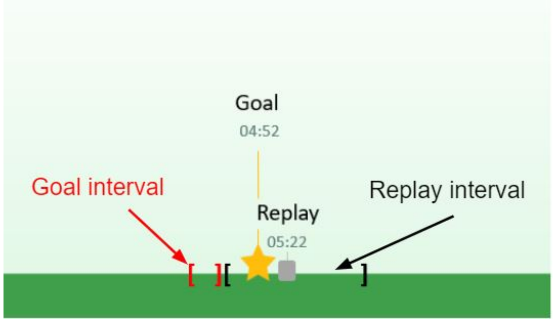
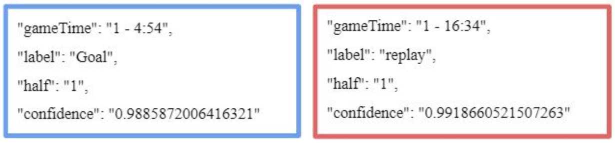


Fig. 3: Run Time When Two Highlights Overlap

D. Post Processing

From the extracted features, we got 4690 labels using the model trained by the SoccerNet-v2 training set. When the threshold was set to 0.9 for confidence for “Goal”, all of the goals in the first half were found. When actually producing the highlight video, we took instances where the label was “Goal” and the confidence was more than or equal to 0.5 and sorted them. If the number of those instances exceeded $n+2$, where n is the actual number of goals in the original video, we only chose the $n+2$ instances with the best confidence as highlights for the resulting video. If it did not, no additional selection process was made for instances marked ”Goal”. For ”replay”, all instances with a confidence of 0.95 and above were included in the highlight.

The maximum difference between the exact and the predicted goal time was 6 seconds. This is encouraging considering that the mAP score is calculated based on a time difference of 5 seconds. Since there could be a 5-second difference, the selected highlights started 10 seconds before the label time and ended 10 seconds after (See Fig. 2) If two highlights overlapped, we set the run time as shown in Fig. 3. From this post-processing method, we were able to produce the highlight video.



```
"gameTime": "1 - 4:54",
"label": "Goal",
"half": "1",
"confidence": "0.9885872006416321"
```

```
"gameTime": "1 - 16:34",
"label": "replay",
"half": "1",
"confidence": "0.9918660521507263"
```

Fig. 4: Annotations of Action (Left) and Highlight Classification (Right)

IV. Experiments

A. Dataset

The dataset we used is SoccerNet-v2, which has extended annotations on the SoccerNet dataset [7], [19]. SoccerNet is a large-scale dataset for soccer video understanding. The dataset contains 500 videos of completely broadcasted soccer games from six major European championships from 2014 to 2017. Each game has two 45-minute videos divided into the first half and second half. The first annotations in SoccerNet-v1 [15] cover temporal timestamps of three main actions in soccer: goals, cards, and substitutions. SoccerNet-v2 has increased the number of annotations to 17 actions including penalties, clearances, balls out of play, and so on. The single time step at which each action starts is annotated in the order of the game time. The dataset also contains feature vectors for each video at 2 frames per second. The feature vectors were extracted from the pretrained ResNet152 and reduced into 512 dimensions to be used as training data instead of video frame images.

SoccerNet dataset also contains annotations on the timestep of replay scenes which we used as a highlight scene in this paper. As shown in the right side of Fig. 4, each label data includes the start time of the camera angle changes, which is denoted among the three labels: live, replay, and others. We changed it into a binary highlight classification task by defining the replay label as a highlight class and defining live and others as a non-highlight class.

B. Evaluation Metrics

As an evaluation metric, we used the average-mAP which is the Mean Average Precision from the PR curve that is averaged over the classes. However, since it is impossible to predict the exact single timestep of the start of the action or highlight classes, we used mAP with the tolerance value. The tolerance value represents how our prediction is close to the ground truth label. A predicted action class is positive if the gap between the predicted and ground truth time step falls within a given tolerance. The tolerance value was tested from 5s to 60s with a step size of 5s. For action classification, we computed the average-mAP through all 17 action classes. For highlight classification, we computed for the two labels which are replay and live or others.

C. Highlight Classification

As we defined our highlight class as the replay of the important actions, we trained our binary classification models which classify each frame image as a replay or not. NetVLAD++ model is selected as a backbone model in this task since it showed the highest performance in the action classification task than the other baseline models. For the feature extraction model, we fixed the weights of pretrained NetVLAD++ as a feature extractor and trained only the final MLP classifiers. Otherwise, in the fine-tuning method, the number of layers and dimensions of the last classifier has changed, and then the model is trained again with new highlight labels. The performance of the highlight classification is computed using the replay label of the closest time for every video chunk.

TABLE I: Average-mAP Results of Action Spotting Baseline Models

Action Class	CALF	NetVLAD++
All	0.389	0.510
Penalty	0.373	0.665
Kick-off	0.327	0.603
Goal	***0.682***	***0.704***
Substitution	0.425	0.704
Offside	0.241	0.370
Shots on target	0.249	0.377
Shots off target	0.296	0.380
Clearance	0.496	0.563
Ball out of play	0.630	0.685
Throw-in	0.555	0.654
Foul	0.510	0.627
Indirect free-kick	0.361	0.438
Direct free-kick	0.406	0.560
Corner	0.693	0.790
Yellow card	0.356	0.522
Red card	0.010	0.014
Yellow → Red card	0.001	0.011

For highlight classification, we compared feature extraction and fine-tuning methods, and Table II shows the results for our experiments. We tested for single and multiple MLP classifiers for both feature extraction and fine-tuning methods. For classifiers with two MLP layers, the number of hidden dimensions is compared for 512, 1024, and 2048 in order to select the model with the best performance. In the result, the fine-tuning model had better performance in average-mAP than that of models with extracted features. Since we changed our target label from action to highlights, the weights for NetVLAD++ should be optimized again. Also, the performance was best when the number of MLP layers of the model was 2 and the hidden dimension of the MLP was 1024. This result represents that the extracted features might not have enough information to be optimized with the classifier of 2048-dimensions. Also, the average-mAP of the replay class is consistently higher than the live class. This might be because the beginning of the playback scene is fixed with some logos during the broadcast and the model could capture that point well.

D. Highlight Generation of Real Soccer Videos

Other than the highlight classification, we combined the result of the action classifier and highlight classifier to generate the final highlight videos of the full soccer game. Since the ground truth highlight video does not exist, it was not available to test our generated highlight videos quantitatively. Hence, we tested our generation model with the existing soccer game broadcasts of 2022 WorldCup using the model trained with the

TABLE II: Average mAP for Highlight Classification

Model Structure	Number of MLP Layers	Hidden Dimension	All	Live or Others	Replay
Feature Extraction	1	-	0.523	**0.3787**	0.6672
	2	512	0.5204	0.3743	0.6665
	2	1024	0.5225	0.3725	0.6725
	2	2048	0.5215	0.3709	0.672
Fine Tuning	1	-	0.5254	0.3781	0.6727
	2	512	0.5228	0.3733	0.6724
	2	1024	**0.5286**	**0.3787**	**0.6786**
	2	2048	0.5232	0.3721	0.6742

SoccerNet dataset. The code and generated highlight videos are available in our github link [1]

V. Conclusion

Throughout this paper, we proposed a deep-learning-based automatic highlight generator for any kind of soccer video. To generate natural highlight videos, we defined replayed parts of the original video as a highlight. In addition, we designed our model to include all goal scenes, as we took people's common stereotypes about highlights into account. Therefore, we utilized NetVLAD++ as a goal classifier and as a backbone model for a highlight classifier. By merging these two kinds of classifiers, we proposed a natural automatic highlight generator for soccer games. Our result can be improved by using metadata combined with audio data, adding more refined video data, or changing the structure of the model. Hence, it will be a springboard for automatic video highlight generation systems.

Acknowledgment

This research was supported by the Ministry of Science, ICT(MIST), Korea, under the Project-based AI Talent Fostering Program (RS-2022-00143911) supervised by the Institute for Information & Communications Technology Planning & Evaluation (IITP).

References

[1] E. Apostolidis, E. Adamantidou, A. I. Metsai, V. Mezaris, and I. Patras, "Ac-sum-gan: Connecting actor-critic and generative adversarial networks for unsupervised video summarization," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 8, pp. 3278–3292, 2020.

[2] W. Zhu, Y. Han, J. Lu, and J. Zhou, "Relational reasoning over spatial-temporal graphs for video summarization," *IEEE Transactions on Image Processing*, vol. 31, pp. 3017–3031, 2022.

[3] Y. Jung, D. Cho, D. Kim, S. Woo, and I. S. Kweon, "Discriminative feature learning for unsupervised video summarization," in *Proceedings of the AAAI Conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 8537–8544.

[4] J. A. Ghauri, S. Hakimov, and R. Ewerth, "Supervised video summarization via multiple feature sets with parallel attention," in *2021 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2021, pp. 1–6s.

[5] E. Apostolidis, G. Balaouras, V. Mezaris, and I. Patras, "Combining global and local attention with positional encoding for video summarization," in *2021 IEEE International Symposium on Multimedia (ISM)*. IEEE, 2021, pp. 226–234.

[6] ——, "Summarizing videos using concentrated attention and considering the uniqueness and diversity of the video frames," in *Proceedings of the 2022 International Conference on Multimedia Retrieval*, 2022, pp. 407–415.

[7] A. Deliege, A. Cioppa, S. Giancola, M. J. Seikavandi, J. V. Dueholm, K. Nasrollahi, B. Ghanem, T. B. Moeslund, and M. Van Droogenbroeck, "Soccernet-v2: A dataset and benchmarks for holistic understanding of broadcast soccer videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 4508–4519.

[8] L. Wang, D. Q. Huynh, and P. Koniusz, "A comparative review of recent kinect-based action recognition algorithms," *IEEE Transactions on Image Processing*, vol. 29, pp. 15–28, 2019.

[9] H. Xu, A. Das, and K. Saenko, "R-c3d: Region convolutional 3d network for temporal activity detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 5783–5792.

[10] S. Karaman, L. Seidenari, and A. Del Bimbo, "Fast saliency based pooling of fisher encoded dense trajectories," in *ECCV THUMOS Workshop*, vol. 1, no. 2, 2014, p. 5.

[11] Z. Shou, D. Wang, and S.-F. Chang, "Temporal action localization in untrimmed videos via multi-stage cnns," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1049–1058.

[12] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2018, pp. 6450–6459.

[13] G. Bertasius, H. Wang, and L. Torresani, "Is space-time attention all you need for video understanding?" in *ICML*, vol. 2, no. 3, 2021, p. 4.

[14] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "Slowfast networks for video recognition," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 6202–6211.

[15] S. Giancola, M. Amine, T. Dghaily, and B. Ghanem, "Soccernet: A scalable dataset for action spotting in soccer videos," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2018, pp. 1711–1721.

[16] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "Netvlad: Cnn architecture for weakly supervised place recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 5297–5307.

[17] A. Cioppa, A. Deliege, S. Giancola, B. Ghanem, M. V. Droogenbroeck, R. Gade, and T. B. Moeslund, "A context-aware loss function for action spotting in soccer videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 13 126–13 136.

[18] S. Giancola and B. Ghanem, "Temporally-aware feature pooling for action spotting in soccer broadcasts," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 4490–4499.

[19] S. Giancola, A. Cioppa, A. Deliège, F. Magera, V. Somers, L. Kang, X. Zhou, O. Barnich, C. De Vleeschouwer, A. Alahi *et al.*, "Soccernet 2022 challenges results," in *Proceedings of the 5th International ACM Workshop on Multimedia Content Analysis in Sports*, 2022, pp. 75–86.

[20] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[21] Z. Li and D. Hoiem, "Learning without forgetting," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 12, pp. 2935–2947, 2017.

[1] https://github.com/JwaYounkyung/LSMMA_Project.git