Proceedings of 12th IOE Graduate Conference
Peer Reviewed
ISSN: 2350-8914 (Online), 2350-8906 (Print)
Year: 2022 Month: October Volume: 12

Assisting Soccer Game Summarization via Audio Intensity Analysis of Game Highlights

Sushant Gautam [a], Cise Midoglu [b], Saeed Shafiee Sabet [c], Dinesh Baniya Kshatri [d], Pål Halvorsen [e]

[a, d] *Thapathali Campus, Institute of Engineering, Tribhuvan University, Nepal*
[b, c, e] *Simula Metropolitan Center for Digital Engineering (SimulaMet), Norway*
[e] *OsloMet, Norway*
✉ [a] sushant76msiise20@tcioe.edu.np, [b] cise@simula.no, [c] saeed@simula.no, [d] dinesh@ioe.edu.np, [e] paalh@simula.no

Abstract

In association football, the development of multimodal summaries is of great importance to both broadcasters and spectators since a large number of viewers choose to follow just the soccer game highlights. The fundamental drive for the development of summarization systems is the requirement to manage huge amounts of data in different formats. By highlighting the most pertinent facts and limiting or omitting unnecessary aspects, summarization helps avoid ”information overload.” The properties of the audio signals during a particular event can be used to calculate excitement among the audience when the event happened and filter events based on their importance to the audience. A root-mean-square (RMS) analysis of audio events was carried out to analyse the excitement across the events in the SoccerNet dataset. It was clearly seen that important events with excitement have a high and distinguishable RMS audio intensity. It was also observed that the generated noise of the crowd was significantly different across various events between the home and away team. The intensity was higher for events related to the home team. Likewise, as the wavelet has the benefit of integrating a wave with a specific period, Morlet wavelet analysis was performed for various event types, and the power of the signal across various wavelet scales was analyzed. A distinct signature across various wavelet scales was observed for different events.

Keywords

association football, audio signal, soccer game highlights, summarization

1. Introduction

Soccer’s market position in the global sports industry is enormous, and the demand for soccer-related video material continues to climb[1]. In this sense, it is vital to provide game summaries and an overview of the game’s most critical events. In spite of this, annotating and creating events and summaries involves expensive equipment and a substantial amount of tedious, arduous human labour. The automation of the soccer game summary and perhaps summarization process allows for the speedy production of game highlights at a much-reduced price.

Due to the fact that a large number of viewers opt to watch just the game’s biggest highlights, the creation of multimodal summaries is crucial for both broadcasters and fans[2]. The availability of a huge number of games and multimedia material heightens the need of real-time or near-real-time summarization systems[3]. The need to handle vast quantities of data in multiple forms is the driving force behind the creation of summarization systems. By emphasising the most important information and reducing or deleting irrelevant details, a summary helps prevent ”information overload.”

Our contributions are as follows:

- We explored RMS-based and wavelet-based audio intensity analysis methods to evaluate excitements on audience.
- We analysed audio in the game videos in SoccerNet[4] dataset and found that interesting events (goal, penalty, etc) have higher audio intensity.

The rest of this paper is organized as follows. In

Section 2, the background information and a brief overview of related work are provided. The SoccerNet dataset to be used for preliminary analysis is elaborated in Section 4. In Section 3, the proposed approach is described in detail, including various alternative methods for audio intensity analysis. Demonstration of the preliminary findings from the SoccerNet dataset is presented in Section 5. The initial results and insights are discussed in Section 5.3. Section 6 concludes the paper.

2. Background and Related Work

2.1 Sports Analytics and Related Applications

Whilst video material has been the primary source for automated knowledge extraction in soccer for many years, event data offers richer context information and needs less processing, making it more relevant in sports analytics [2]. Nanayakkara et al. [5] explored the importance of text-based commentaries available from sports websites, along with temporal information, for event classification in sports analytics.

2.2 Soccer Game Summarization

The demand for soccer game summarization is increasing due to the preference of a large audience to view only important parts of events in video form as well as textual summaries of the game. Proper event detection and highlight prioritization are absolutely essential for a good summarization system[6, 7, 3]. A general framework for sports video summarization and its application to soccer footage is provided by earlier work by Li et al. [8]. An extractive multi-modal summarization (MMS) technique is provided by Li et al. [9] for asynchronous collections of text, images, audio, and video to generate a textual summary.

In their summarising process, Sanabria [2] combines event and audio data as input to a hierarchical recurrent neural network to determine which recommendations should be included in the summary. They utilise audio to assess the game's level of excitement.

2.3 Event Detection and Highlights Prioritization

Event selection and prioritization in the summarization system are usually done with static rules for different types of events. Earlier works utilised audio intensity information along with heuristic criteria derived from human analysis of sports video footage to predict goal events in basketball videos [10].

Hanjalic et al. [11] used the variations in sound energy across a video as one of the excitement components, which are then used to search for highlights at points when the video's content elicits tremendous enthusiasm in the viewer. Tjondronegoro et al. [12] employed audio for a summarization technique, recognising whistle sounds dependent on the audio's frequency and pitch. Kime et al. [13] proposed an automated method for extracting goal events from soccer videos based only on audio characteristics, without the use of costly video-based features. Duxans et al. [14] used acoustic aspects of the audio track, namely the block energy and the acoustic repetition index, for identifying and creating video summaries of soccer game highlights. Recent work by Rongved et al. [15] evaluates different neural networks based approaches with model fusion on different combinations with audio and video, to identify and categorise events in soccer footage using the SoccerNet dataset. Combining additional information from game metadata and captions can aid the event Detection and highlights the prioritization process.

2.4 MultiModel Approaches

Raventos et al. [16] condemn methodologies that focus only on visual analysis and overlook the vital information offered by the audio channel. It emphasizes the importance of audio detectors, which are essential for the interpretation and scoring of video images. Using audio-visual Convolutional Neural Network (CNN) characteristics, Haruyama et al. [17] suggested a method for evaluating noteworthy soccer video sequences using weighted majority voting based on SVM-based calculations. The results indicate that evaluating both auditory and visual sequences is advantageous for understanding the importance of critical moments, particularly in soccer games where game situations impact not only player actions but also the applause of the crowd. Vanderplaetse et al. [18] analysed a variety of methods for incorporating audio stream into video-only-based architectures for 500 game videos in SoccerNet dataset [4]. Significant improvements were found with such fusion for both the action classification and the action spotting task over a video-only baseline. Recently Gautam et al. [19] presented a soccer game summarization pipeline that could generate a summary from the input captions or metadata of the events. They use audio levels in the video around the event timeframe to calculate the intensity that was used to filter the events to be included in the summary. Figure 4 shows Audio Viz Dashboard used in Gautam et al.[19]. The RMS audio intensity around the event is used to rank the event. Important events, according to the excitement in the audio, are filtered to be included in the summary.

3. Proposed Methodology

Audio intensity analysis can support automatic video summarization pipelines, by enabling the calculation of excitement around different events, and can be used as a filter to select game highlights based on audience-perceived importance. In this section, we propose alternative approaches for audio intensity analysis.

Figure 1 shows the use of audio filter in summarization pipeline. The audio from a particular event is used to calculate excitement around that event and filter the events based on their importance.

Root-mean-square (RMS) Analysis Using the audio samples or spectrogram, the RMS value for each audio frame may be determined. The RMS value of a signal is determined as the square root of the average of the signal sample's squared values[20]. For a collection of complex-valued signals represented as N discrete sampled values $-[x_0, x_1, \cdots, x_{N-1}]$, the mean square value x_{RMS} is provided as in Equation 1.

$$x_{RMS} = \sqrt{\frac{|x_0|^2 + |x_1|^2 + \cdots + |x_{N-1}|^2}{N}} = \sqrt{\frac{1}{N}\sum_{n=0}^{N-1} |x[n]|^2} \tag{1}$$

RMS value calculation from audio samples seems to be efficient since no STFT computation is required. However, since spectrogram frames can be windowed, spectrograms provide a more precise depiction of energy across time; hence, the use of spectrogram for RMS value calculation is preferred. Using frequency domain components $X[k]$ and Parseval's theorem, it is also possible to calculate the root mean square value as in Equation 2.

$$x_{RMS} : \sqrt{\frac{1}{N}\sum_{n=0}^{N-1} |x[n]|^2} = \sqrt{\frac{1}{N^2}\sum_{k=0}^{N-1} |X[k]|^2} \tag{2}$$

Wavelet Analysis The wavelet has the benefit of combining a wave with a certain period and being limited in size. The Morlet wavelet, a popular and straightforward wavelet, as seen in Figure 2, is actually a Sine wave multiplied by a Gaussian envelope.

The formula for mother wavelet for the Morlet wavelet transform is shown in Equation 3.

$$\psi_0(\eta) = \pi^{-1/4} e^{i\omega_0\eta} e^{-\eta^2/2} \tag{3}$$

The number of oscillations inside the wavelet itself is denoted by the wavenumber w_0. In practice, $w_0 = 6$ is set such that the wavelet's mean is zero. To effectively sample all frequencies included in the time series, a set of scaling parameters s, also known as the scale or period, is selected. First, the lowest resolvable scale, $s0$, is selected and then multiplied by a constant multiple up to the maximum scale. The greatest scale should be less than half of the whole time series.

4. Dataset and Dashboard

SoccerNet [4] is an open dataset that focuses on the localization of rare occurrences in extensive video game footage. It comprises a total of 764 hours of 500 uncut soccer broadcasts, annotated with 3 key event categories: *goal*, *yellow/red card*, and *player substitution*. The annotation granularity is adjusted to a one-second resolution.

Listing 1: Data structure for the action spotting task of SoccerNet-V2. (*) Field optional.

```
{
#  Start
      "UrlLocal"          :  <path>,
      "UrlYoutube"(*)     :  <link>,
      "annotations"       : [{  ...  }],
#  End
      "gameAwayTeam"      :  <team name>,
      "gameDate"          :  <timestamp>,
      "gameHomeTeam"      :  <team name>,
      "gameScore"         :  <int> - <int>
}
```

SoccerNet-V2 [1] extends the number of action classes from 3 to 17. The structure of the dataset for the SoccerNet-V2 action detection task is shown in Listing 1, along with the structure of the event annotations in 2. Figure 3 depicts the per-game distribution of the occurrence of different types of events in the SoccerNet dataset.

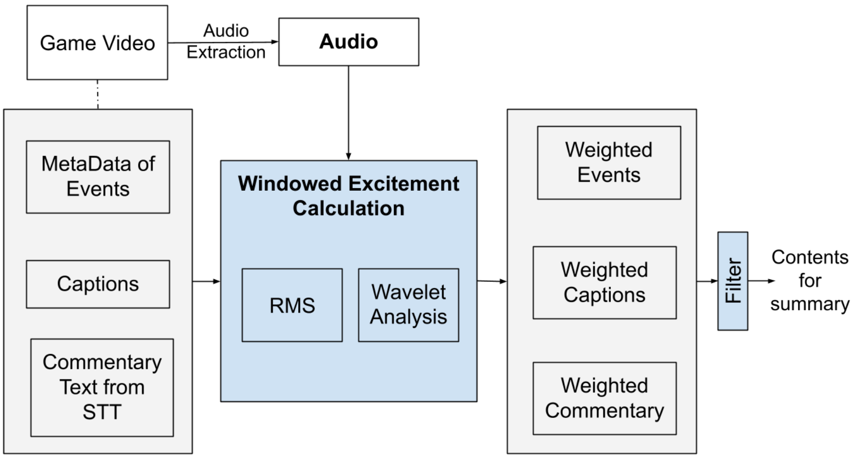


Figure 1: Audio-based Filter in Summarization Pipeline

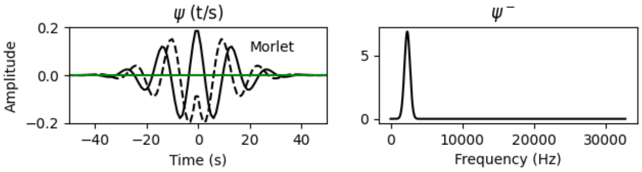


Figure 2: Mother Morlet Wavelet

Listing 2: Annotation structure for sample event in the action spotting task of SoccerNet-V2.

```
{
    ”gameTime”: <half Number> − <time f
    ”label”       : <action type>,
    ”position”
: <time in ms from beginning of game>,
    ”team”
: <home/away/not applicable>,
    ”visibility”: <visible/not shown>
}
```

In this work, we use the SoccerNet-V2 dataset for the action spotting task, to demonstrate various audio intensity analysis methods. We also make use of ”AudioViz”, an audio intensity dashboard implemented by [19] for understanding the correlation between audio intensity levels and the events in a soccer game video. This dashboard plays the game video along with indicators for the corresponding audio levels, event annotations, and an ordered list of the top events in the game during which the audio intensity was highest. The dashboard can be used as a validator for filtering game highlights according to perceived importance (based on audio intensity). The tool is provided as open-source software for the community.

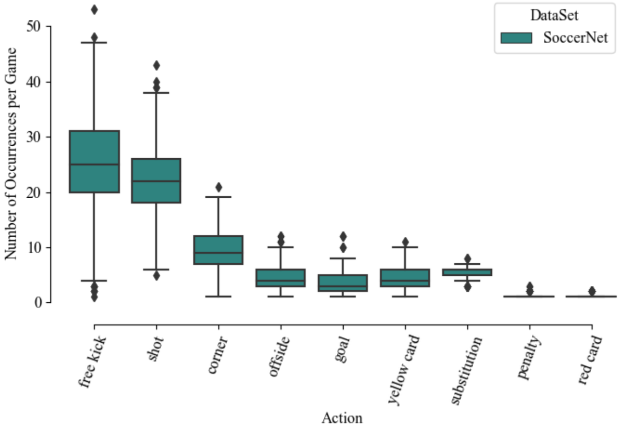


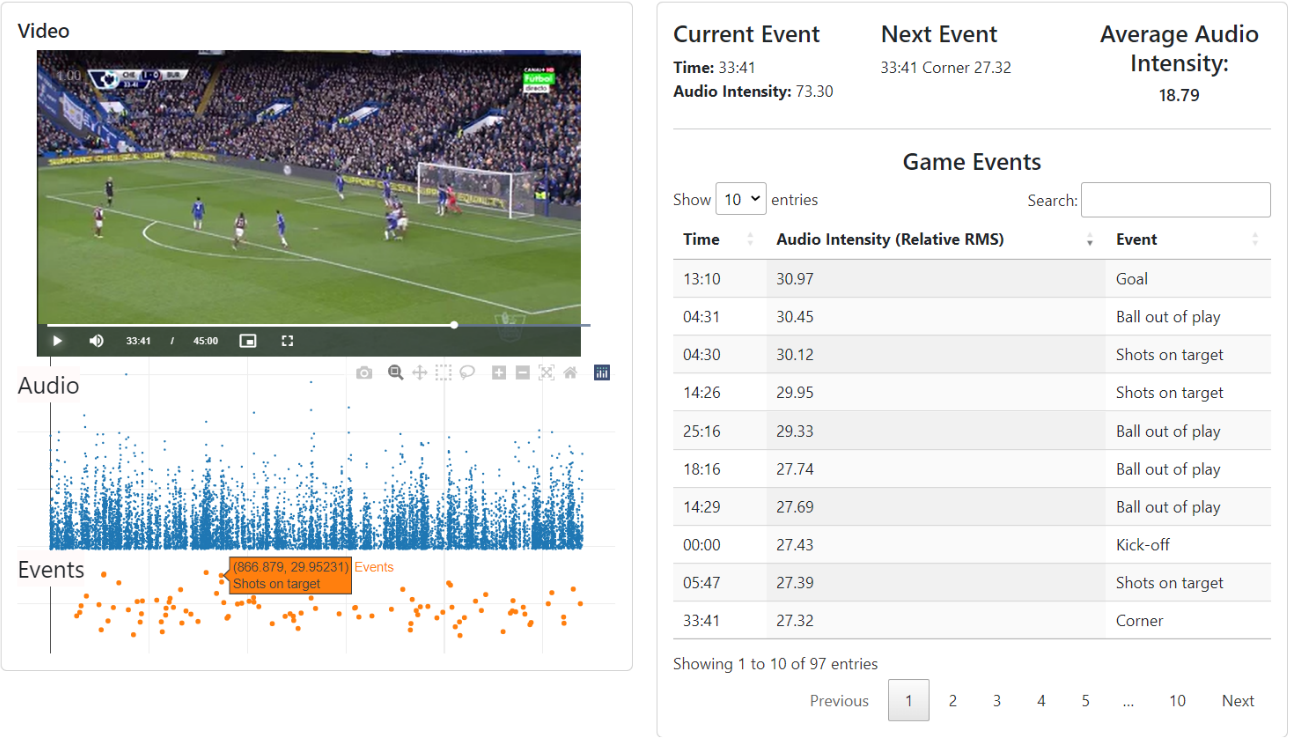
Figure 3: The distribution of the number of occurrences of different types of events per game in SoccerNet.

5. Experiments and Results

In this section, we try to answer a number of research questions using the proposed methodology outlined in Sections 3 and 4.

1. Is there a significant difference between different types of events in terms of the overall intensity of the game audio around the event (including all artefacts such as audience cheer, commentator voice, etc.)? If so, what is the order of the magnitude of the overall audio intensity for each type of event?

2. Is there a difference between the overall audio intensity associated with events of the same type,



Video

33:41 / 45:00

Audio

Events

(866.879, 29.95231) Events
Shots on target

Current Event

Time: 33:41
Audio Intensity: 73.30

Next Event

33:41 Corner 27.32

Average Audio Intensity:

18.79

Game Events

Show 10 entries

Search:

Time	Audio Intensity (Relative RMS)	Event
13:10	30.97	Goal
04:31	30.45	Ball out of play
04:30	30.12	Shots on target
14:26	29.95	Shots on target
25:16	29.33	Ball out of play
18:16	27.74	Ball out of play
14:29	27.69	Ball out of play
00:00	27.43	Kick-off
05:47	27.39	Shots on target
33:41	27.32	Corner

Showing 1 to 10 of 97 entries

Previous 1 2 3 4 5 ... 10 Next

Figure 4: AudioViz Dashboard

according to whether the event is related to the home team or the away team?

3. Is there a difference between the overall audio intensity associated with events, according to whether the event happened in the first half or the second half of the game?

4. Is there a difference between the audio associated with different types of events, in terms of frequency composition?

5.1 Root-mean-square(RMS) Analysis

Using this method, we try to answer questions 1, 2 and 3.

Source	df	F	Sig.
Event	15	666.529	.000
Host	1	6.614	.010
HalfTime	1	.004	.949
Event *Host	15	17.506	.000
Event * HalfTime	15	2.514	.001
Host * HalfTime	1	.249	.617

a. R Squared = .096 (Adjusted R Squared = .095)

Table 1: Tests of Between-Subjects Effects for Dependent Variable: Audio RMS

To answer the research questions a generalized linear

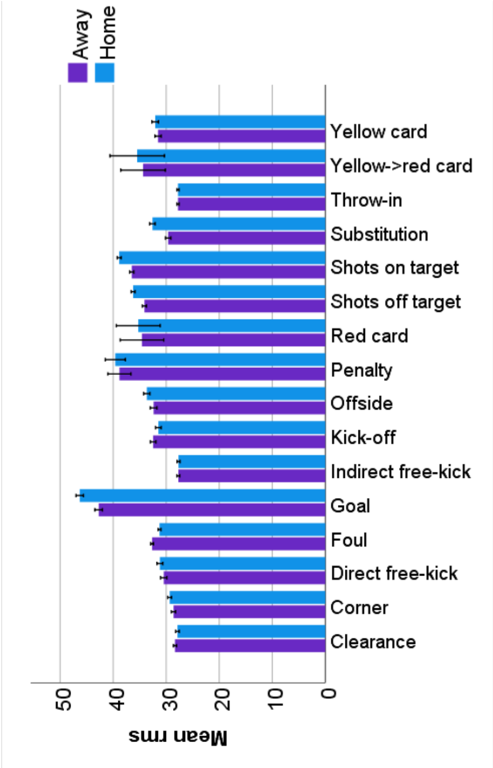


Figure 5: Mean audio RMS between different events related to home and away team.

model (GLM) was conducted to compare the RMS in different events and home vs away matches. The results are reported in Table 1, where *Event* indicates 16 different types of events in the dataset, *Host* indicates whether the event was from a home or away team, and *HalfTime* indicates whether the event occurred in the first or second half. The results indicate that there is a significant main effect of the event on the RMS, and a significant main effect on home/away matches. This means that as we expected the generated noise of the crowd was significantly different across various events and if it happened for the home or the away team. There was no significant main effect found for half, meaning that the level of noise from the crowd stayed the same on different half times. However, there is a significant interaction between half and type of event, meaning that there some events create a different noise across two different half times. In addition, a significant interaction found between the hosting and the event, meaning that depending on whether it's home or away, events have a different noise, for example, goals for the home team generated a significantly higher RMS than away matches as shown in 5.

Figure 6 shows the distribution of audio RMS of different types of events. The higher RMS value represents the higher excitement in the game. The order of importance of events across the dataset can be seen in Figure 6. It was found that the goal event has the highest RMS intensity followed by the penalty event.

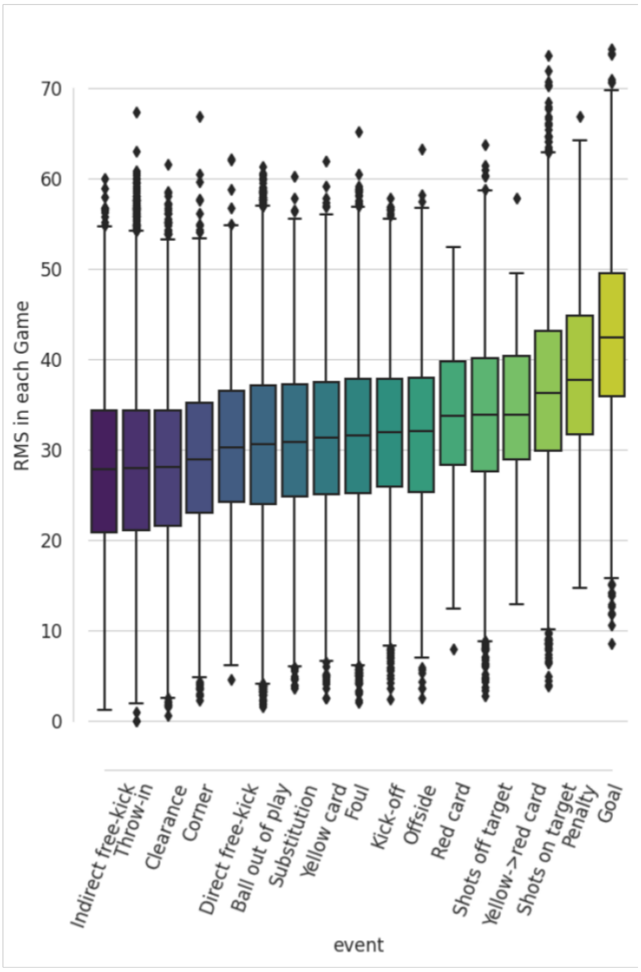


Figure 6: The distribution of audio RMS of different types of events.

5.2 Wavelet Analysis

Using this method, we try to answer question 4.

Figure 7 shows a time-series of 1-second audio sample from a goal event form the dataset. Figure 8 shows the normalized wavelet power spectrum for the goal event with an audio time-series shown in Figure 7. The y-axis shows the variation of period, which ranges from the smallest resolvable scale of 2 to 16384, which is around half of the full-time series samples. The hatched area has also been termed the cone of influence, with its conical boundary being the contour line for significance level.

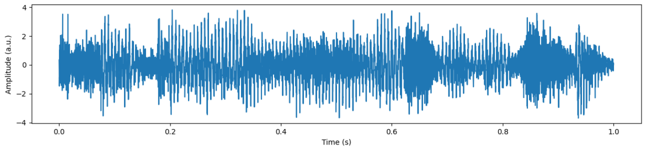


Figure 7: 1-second time audio series after a goal event

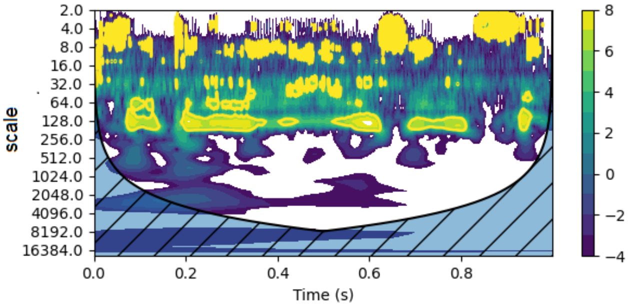


Figure 8: Wavelet Power Spectrum

Figure 9 shows the global wavelet spectrum for the goal event with the wavelet power spectrum shown in Figure 8. The gray level indicated the Fourier spectrum. The blue line in the figure represents the global wavelet spectrum, with the red line reflecting

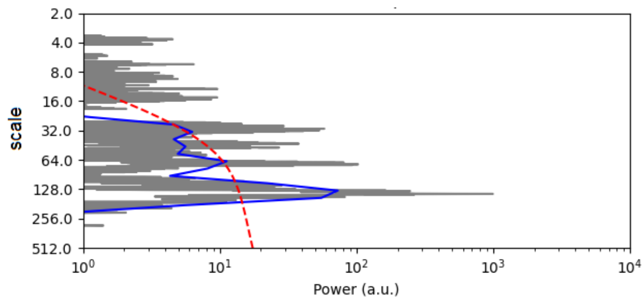


Figure 9: Global Wavelet Spectrum

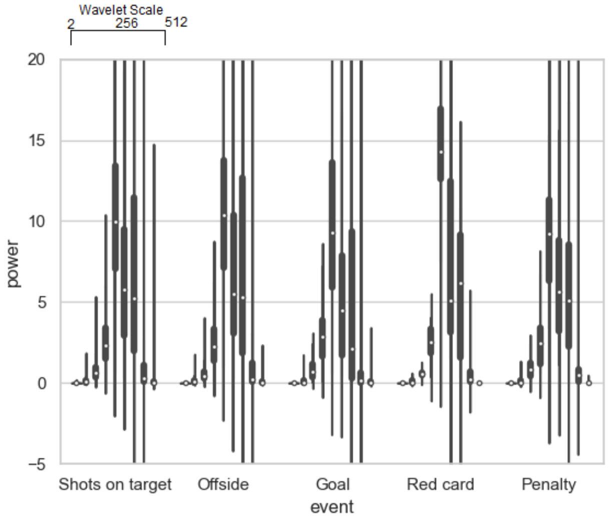


Figure 10: Power distribution across Wavelet scales

the significance level. The y-axis only shows the variation of period from the smallest resolvable scale of 2 to 512, as the results around the higher order were not much interesting, as seen in Figure 8.

Figure 10 shows a multiple-box-plot of power values like shown in global wavelet spectrum in Figure 8.

Only periods from 2 to 512, which are multiples of 2, are chosen for only five of the different events. A distinct signature across various wavelet scales was observed for different events. For instance, the power of scale 32 was distinctly higher for red-card events when compared to other event types. The power at scales 32 and 128 was seen to be highest in shots, offsides, and goal events.

5.3 Discussion

This work explored the use of filters using audio as input to the summarization pipeline to decide which events should be in the summary. According to the RMS analysis, it was found from the analysis of the audio from the videos that interesting events have high audio intensity and they capture the excitement of the crowd. It is seen from Figure 6 that the goal event had the largest distribution of RMS audio intensity, making it the most intriguing aspect of the game for the audience. Similarly, additional noteworthy occurrences include penalties, shots, and card events, in that order. Similarly, RMS analysis demonstrated that the acoustic intensity of home team events is greater than that of away team events. However, there was no notable difference between the first and second half in terms of the intensity of the incidents. Wavelet analysis showed that the distribution of the power at different scales was different across events. The distribution of the power at different scales of 'Shots on target' events was found to be similar to that of 'goal event'. Interestingly, the power at scales 32 was the highest among other scales across all the events.

6. Conclusion

The use of audio as input to the filters in the summarization pipeline to capture the importance of the event was explored. Since it was clearly seen that important events with excitement have high and distinguishable audio intensity, they are good candidates to be used in the filter module of a summarization pipeline. The exploration also showed that the events for the home team have higher excitement compared with the events for the away team. However, the statistics showed that the difference between the intensities between different halves was not very significant in the dataset explored.

Acknowledgment

This work was carried under the Norwegian Research Council, project number 327717 (AI-producer).

References

[1] Adrien Deliege, Anthony Cioppa, Silvio Giancola, Meisam J. Seikavandi, Jacob V. Dueholm, Kamal Nasrollahi, Bernard Ghanem, Thomas B. Moeslund, and Marc 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 (CVPR) Workshops*, pages 4508–4519, June 2021.

[2] Melissa Sanabria, Sherly, Frédéric Precioso, and Thomas Menguy. A Deep Architecture for

Multimodal Summarization of Soccer Games. In *MMSports '19: Proceedings Proceedings of the 2nd International Workshop on Multimedia Content Analysis in Sports*, pages 16–24. Association for Computing Machinery, New York, NY, USA, October 2019.

[3] Cise Midoglu, Steven A. Hicks, Vajira Thambawita, Tomas Kupka, and Pål Halvorsen. Mmsys'22 grand challenge on ai-based video production for soccer. In *13th ACM Multimedia Systems Conference (MMSys'22)*. ACM, 2022.

[4] Silvio Giancola, Mohieddine Amine, Tarek Dghaily, and Bernard Ghanem. Soccernet: A scalable dataset for action spotting in soccer videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 1711–1721, June 2018.

[5] Nanayakkara Kpk. *Event classification from text based commentaries for sports analytics*. PhD thesis, University of Moratuwa, 2019.

[6] Joakim Olav Valand, Haris Kadragic, Steven Alexander Hicks, Vajira Lasantha Thambawita, Cise Midoglu, Tomas Kupka, Dag Johansen, Michael Alexander Riegler, and Pål Halvorsen. AI-Based Video Clipping of Soccer Events. *Mach. Learn. Knowl. Extr.*, 3(4):990–1008, December 2021.

[7] Abdullah Aman Khan, Yunbo Rao, and Jie Shao. ENet: event based highlight generation network for broadcast sports videos. *Multimedia Systems*, pages 1–12, July 2022.

[8] Baoxin Li, Hao Pan, and Ibrahim Sezan. A general framework for sports video summarization with its application to soccer. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, volume 3, pages III–169. IEEE, 2003.

[9] Haoran Li, Junnan Zhu, Cong Ma, Jiajun Zhang, and Chengqing Zong. Read, Watch, Listen, and Summarize: Multi-Modal Summarization for Asynchronous Text, Image, Audio and Video. *IEEE Trans. Knowl. Data Eng.*, 31(5):996–1009, June 2018.

[10] Surya Nepal, Uma Srinivasan, and Graham Reynolds. Automatic detection of 'Goal' segments in basketball videos. In *MULTIMEDIA '01: Proceedings of the ninth ACM international conference on Multimedia*, pages 261–269. Association for Computing Machinery, New York, NY, USA, October 2001.

[11] A. Hanjalic. Generic approach to highlights extraction from a sport video. In *Proceedings 2003 International Conference on Image Processing (Cat. No.03CH37429)*, volume 1, pages I–1. IEEE, September 2003.

[12] Dian Tjondronegoro, Yi-Ping Phoebe Chen, and Binh Pham. Sports video summarization using highlights and play-breaks. In *Proceedings of ACM SIGMM International Workshop on Multimedia Information Retrieval (MIR)*, page 201–208, 2003.

[13] Hyoung-Gook Kim, Steffen Roeber, Amjad Samour, and Thomas Sikora. Detection of goal events in soccer videos. *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, 5682:317–325, January 2005.

[14] Helenca Duxans, Xavier Anguera, and David Conejero. Audio based soccer game summarization. In *2009 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting*, pages 1–6. IEEE, May 2009.

[15] Olav Andre Nergård Rongved, Markus Stige, Steven Alexander Hicks, Vajira Lasantha Thambawita, Cise Midoglu, Evi Zouganeli, Dag Johansen, Michael Alexander Riegler, and Pål Halvorsen. Automated Event Detection and Classification in Soccer: The Potential of Using Multiple Modalities. *Mach. Learn. Knowl. Extr.*, 3(4):1030–1054, December 2021.

[16] Arnau Raventós, Raúl Quijada, Luis Torres, Francesc Tarrés, Eusebio Carasusán, and Daniel Giribet. The importance of audio descriptors in automatic soccer highlights generation. In *2014 IEEE 11th International Multi-Conference on Systems, Signals & Devices (SSD14)*, pages 1–6. IEEE, February 2014.

[17] Tomoki Haruyama, Sho Takahashi, Takahiro Ogawa, and Miki Haseyama. Estimation of Important Scenes in Soccer Videos Based on Collaborative Use of Audio-Visual CNN Features. In *2018 IEEE 7th Global Conference on Consumer Electronics (GCCE)*, pages 710–711. IEEE, October 2018.

[18] Bastien Vanderplaetse and Stéphane Dupont. Improved soccer action spotting using both audio and video streams. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 3921–3931, 2020.

[19] Sushant Gautam, Cise Midoglu, Saeed Shafiee Sabet, Dinesh Baniya Kshatri, and Pål Halvorsen. Soccer Game Summarization using Audio Commentary, Metadata, and Captions. In *NarSUM '22: Proceedings of the 1st Workshop on User-centric Narrative Summarization of Long Videos*, pages 13–22. Association for Computing Machinery, New York, NY, USA, October 2022.

[20] Brian McFee, Colin Raffel, Dawen Liang, Daniel P. W. Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. librosa: Audio and Music Signal Analysis in Python. *Proceedings of the 14th Python in Science Conference*, pages 18–24, 2015.