Soccer Game Summarization using Audio Commentary, Metadata, and Captions

Sushant Gautam
Tribhuvan University, Nepal

Cise Midoglu
SimulaMet, Norway

Saeed Shafiee Sabet
SimulaMet, Norway

Dinesh Baniya Kshatri
Tribhuvan University, Nepal

Pål Halvorsen*
SimulaMet, Norway

ABSTRACT

Soccer is one of the most popular sports globally, and the amount of soccer-related content worldwide, including video footage, audio commentary, team/player statistics, scores, and rankings, is enormous and rapidly growing. Consequently, the generation of multimodal summaries is of tremendous interest for broadcasters and fans alike, as a large percentage of audiences prefer to follow only the main highlights of a game. However, annotating important events and producing summaries often requires expensive equipment and a lot of tedious, cumbersome, manual labour. In this context, recent developments in Artificial Intelligence (AI) have shown great potential. The goal of this work is to create an automated soccer game summarization pipeline using AI. In particular, our focus is on the generation of complete game summaries in continuous text format with length constraints, based on raw game multimedia, as well as readily available game metadata and captions where applicable, using Natural Language Processing (NLP) tools along with heuristics. We curate and extend a number of soccer datasets, implement an end-to-end pipeline for the automatic generation of text summaries, present our preliminary results from the comparative analysis of various summarization methods within this pipeline using different input modalities, and provide a discussion of open challenges in the field of automated game summarization.

CCS CONCEPTS

• **Computing methodologies** → **Artificial intelligence**; **Natural language processing**; • **Information systems** → **Summarization**.

KEYWORDS

AI, football, automated pipeline, NLP, soccer game summary

ACM Reference Format:
Sushant Gautam, Cise Midoglu, Saeed Shafiee Sabet, Dinesh Baniya Kshatri, and Pål Halvorsen. 2022. Soccer Game Summarization using Audio Commentary, Metadata, and Captions. In *Proceedings of the 1st Workshop on User-centric Narrative Summarization of Long Videos (NarSUM '22), Oct. 10, 2022, Lisboa, Portugal.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3552463.3557019

*Also with OsloMet, Norway.

NarSUM '22, October 10, 2022, Lisboa, Portugal

ACM ISBN 978-1-4503-9493-2/22/10.
https://doi.org/10.1145/3552463.3557019

1 INTRODUCTION

Sports broadcasting and streaming are immensely popular, and the interest in viewing videos from sports events grows day by day. Today, live streaming of sports events generates most of the video traffic and is replacing live broadcasting on TV [60]. As of 2020, soccer had a global market share of about 45% of the $500 billion sports industry [54].

The large number of games and the availability of multimedia content make it increasingly important to establish systems for extracting highlights and providing summaries in real-or near real-time. As a large percentage of audiences prefer to follow only the main highlights of a game, the generation of multimodal (video/audio/text) summaries is of tremendous interest to broadcasters and fans alike. The primary impetus for the creation of summarization systems is the need to handle the vast quantities of data available in various formats.

State-of-the-art solutions in this domain are limited, and therefore subject to a lot of interest from researchers, especially in terms of automation. However, unlike action recognition and event detection in videos, event clipping and thumbnail selection for highlight clips, and social media content analysis, there has not been much interest in the journalism or accessibility aspects of video summarization systems. There is a need to establish a framework for getting text and audio summaries of videos automatically, where emphasis on accessibility can prove highly useful for visually impaired audiences as well. Building an automated summarization system may be valuable not only for future games but also for condensing past and completed games.

The overall goal of our work is to implement an automated pipeline which allows for the generation of text and audio summaries of configurable length for a given soccer game, where the inclusion/exclusion of selected important events ("highlights") from the game can be subject to priority configuration input, along with the nature of the generated summaries (neutral vs. taking a particular side, empathic vs. natural, etc.). The pipeline is intended to serve as an automated tool for sports journalists to compile articles with minimal delay upon game completion, as well as to provide global accessibility to soccer game news for visually impaired fans. In this paper, we present a part of our work in progress.

Why generate text summaries? Generating game summaries in the form of continuous text is a comparatively efficient way of distilling the information in comparison to its video counterpart. As languages have expressive advantages, text summaries have multiple applications, ranging from news article generation to social media promotion, as well as increasing accessibility. Using simple

Text to Speech (TTS) technologies, these summaries can also be provided in audio format with minimal effort.

Why not use the game video directly? Computer vision techniques can be used to generate summaries directly from videos through deep video understanding. However, this is a difficult task to accomplish in a single step, and requires systems with a deep understanding of the relationships between different entities in the video. Existing industry practices [4, 9] already create annotations as part of their video production pipeline, which can be exploited for the downstream task of creating game summaries. More lightweight pieces of information such as game audio, metadata, and captions are easier to process than video streams of larger size.

The contributions of this paper are as follows:

- We extend 3 existing soccer datasets, namely SoccerNet [18], SportsSum [25], and K-SportsSum [61], with ground truth summaries, game metadata, and translations, respectively. We provide these publicly as open datasets[1] in order to contribute more resources to the research community for the task of soccer game summarization. We also curate an in-house dataset with a more detailed metadata template, which is a work in progress.
- We design and implement an automated pipeline for generating soccer game summaries in text format using different inputs such as audio commentary, metadata, and captions.
- We present preliminary insights from the comparative analysis of various alternative methods within this pipeline using different input modalities, and compare them against state-of-the-art solutions using objective metrics.
- We provide a discussion of our results as well as open challenges in the field of text summarization within a professional sports context.

The rest of this paper is organized as follows. In Section 2, we provide background information and an overview of related work. In Section 3, we describe our proposed pipeline and various alternative methods for game summarization in detail. In Section 4, we elaborate on our dataset curation. In Section 5, we present our proof-of-concept implementation and preliminary findings. In Section 6, we discuss our initial results, potential applications and open challenges, and derive insights about potential future work. In Section 7, we conclude the paper.

2 BACKGROUND AND RELATED WORK

2.1 Automated Soccer Video Production

State-of-the-art systems for soccer broadcasting and dissemination do not include automated pipelines which can summarize entire soccer games in a configurable and multimodal fashion. The process is dominated by manual efforts, and the focus of automation has overwhelmingly been on the generation of highlight clips through event detection and clipping.

Event detection, also called action detection or action spotting, has received significant attention in the past decade. The goal is to automate the manual event tagging process for soccer videos while maintaining accuracy and efficiency. However, most of the proposed models are computationally expensive and/or relatively inaccurate. Wang et al. [47, 62] proposed TSN, and C3D [55] explored 3D convolution for learning spatio-temporal features. Giancola et al. [19], used ResNet-152 features and NetVLAD pooling to classify events in soccer videos of one-minute length chunks. Rongved et al. [38] showed that the fusion of visual and audio models improves the performance of event detection in the broadcast pipeline.

The amount of work on *event clipping* is more limited. Koumaras et al. [26] presented a shot detection algorithm, and Zawbaa et al. [67] implemented a more tailored algorithm to handle cuts that transitioned gradually over several frames. Zawbaa et al. [66] classified soccer video scenes as long, medium, close-up, and audience/out of field, and several papers presented good results regarding scene classification [42, 64, 66]. As video clips can also contain replays after an event, replay detection can help to filter out irrelevant replay events. Ren et al. [45] introduced the class labels play, focus, replay, and breaks. Detecting replays in soccer games using a logo-based approach was shown to be effective using a Support Vector Machine (SVM) algorithm, but not as effective using an Artificial Neural Network (ANN) [66, 67]. Furthermore, it was shown that audio may be an important modality for finding good clipping points. Raventos et al. [43] used audio features to give an importance score to video highlights, and Tjondronegoro et al. [52] used audio for a summarization method, detecting whistle sounds based on the frequency and pitch of the audio. Finally, some work focused on learning spatio-temporal features using various Machine Learning (ML) approaches [12, 48, 56], and Chen et al. [13] used an entropy-based motion approach to address the problem of video segmentation in sports events. Most recently, Valand et al. [57, 58] exploited logo-transition and scene boundary detection for event clipping using Convolutional Neural Networks (CNNs).

2.2 Video Summarization

Video summarization, also called video abstraction, most commonly refers to the generation of *video* summaries from long videos. Sun et al. [50] offer a pair-wise rating algorithm that learns from web videos and ranks without limitations the highlights of video segments. LiveLight [69] scans and segments the incoming video using a vocabulary. The findings indicate that a saliency-based motion blur filter is the most effective. Li et al. [29] provide a generic framework for sports video summarization and its application to soccer footage. In the first category, keyframes are recovered by treating each pitch as a significant event. For the latter, the broadcaster's audio is evaluated for interesting sections, which are then extracted as keyframes. The framework is shown using a soccer application. Due to the varying length of some games, like cricket, video summaries are difficult to provide. Here, audio data is segmented into brief frames, and the pitch of each frame is calculated to assess the intensity of commentary and crowd applause. Using a decision tree and a set of event rules, boundary, six, wicket, and replay events are classified. Recently, video summarization approaches based on deep learning have been proposed. Outputs may include a keyframe, many keyframes (static storyboards) [22, 30], or a shorter video clip. Gong et al. [20] present a sequential Determinantal Point Process (seqDPP) to describe films in a supervised way, determine the matched frames whose visual distance is below a threshold, and calculate their accuracy, recall, and F-score given two summaries.

[1] https://github.com/simula/soccer-summarization

[37] presents unsupervised video-to-video summarization which is customized by incorporating user feedback. Authors in [35] summarize multiplayer 3D scene games based on game rules. Evaluations of several types of video summaries are presented in [49].

2.3 Text Summarization

There are various works on the generation of text summaries using Natural Language Generation (NLG) systems, in different contexts.

Journalism: In 2007, the Los Angeles Times established a blog dedicated to reporting on murders that would include machine-written text based on a simple template [65]. In 2014, the same journal published the first earthquake-related article using Quakebot [40]. SumTime-Mousam [44] produced weather forecasts with numerical weather prediction data. The system developed by Plachouras et al. [41] could search for financial data using keywords or natural language. For the query "India's GDP 2010", the system locates the record containing India's GDP in 2010 and delivers a text response. An NLG template-based system generated news articles in Finnish, Swedish, and English for the 2017 Finnish municipal elections [28]. Using NLG technology, the BBC was able to upload news articles for each of the United Kingdom's 650 seats on the evening of the 2019 general election [36].

Social media: Graph-based and rate-based methods are the most commonly used for summarizing sports events. Nichols et al. [39] were among the first to consider the possibility of using Twitter for sports event summaries. Their technique is characterized by detecting a sub-event occurrence when the Twitter stream rate exceeds 90% of the previously observed rate. Sharifi et al. [46] used a graph-based method to identify the most frequently occurring terms in a sample of tweets. Then, the system chose a sentence to summarize the incident based on the collection of words that had been discovered. Chua and Asur [15] constructed two Twitter event detection topic models that integrated the temporal aspect of the tweets' terms with the words themselves and picked a group of tweets describing each observed event. Twitinfo [33] recognizes sub-events based on specified keywords using a peak detection algorithm and then gives a timeline-based event display. Kubo et al. [27] addressed the issue of providing sports event summaries in real-time by using Twitter users who are regarded as excellent reporters. The user scores are computed by awarding higher scores to people that post more often during previously discovered subevents inside the event. Hsieh et al. [24] presented a real-time method for analysing soccer matches using their moving-threshold method and the TF-IDF to find the most representative keywords for each sub-event.

2.4 Soccer Game Summarization

In 2008, Chen and Mooney [14] presented a commentary system that describes the events in a given match. It was trained using human-commented games from the Robocup simulation league and introduced three algorithms to generate commentaries for unseen games. It used a probabilistic approach to understand what types of events (e.g., passes or goals) are most likely to be reported on by human commentators. In [31], based on the characteristics of live webcast scripts, methods were developed for sentence extraction and template generation from live webcast scripts. The system extracted significant events during the time period of the live webcast and created a quick summary of the soccer games from the live webcast scripts according to existing rules.

Van der Lee et al. [59] proposed PASS, a template-based data-to-text system that produces two sports reports in Dutch for every game, each with a distinct tone based on the reader's team. The input data is scrapped from online sources, such as the results of past games and previous matches between the teams. The authors contend that the algorithm generates text with a comparable level of variance as GoalGetter [17]. When the scope of a game state is widened or when non-event conversations become prominent on commentaries, even the most advanced state monitoring algorithms suffer. Zhang [68] simplified the formulation such that given the paragraphs of comments on a game at various timestamps, the system detects the occurrence of in-game events. This permits detailed descriptions of states while avoiding the intricacies of several other real-world contexts. K-SportsSum [61] introduced a knowledge-enhanced summarizer that utilizes both live commentaries and the knowledge of sports teams and player to generate sports news.

2.5 Takeaways

Existing soccer game summarization frameworks using Natural Language Processing (NLP) approaches most commonly process sampled frames in a video to identify the objects in the video, and afterwards utilize NLP algorithms.

- The **subtitles** of sports videos, either readily available as detailed **captions** or generated from the audio **commentary** using Speech to Text (STT) systems, are not utilized. The processing of such information, if accessible, and their summarization using NLP algorithms present a potential which we will exploit in this work.
- Similarly, **metadata** from soccer games, which are readily available or can be generated using event detection systems, are infrequently utilized in soccer game summarization.
- Auto-generated text summaries have the problem of being agnostic of certain crucial aspects, particularly in length-restricted scenarios. As a result, variable length text summaries may not ideally describe the complete video to the extent that it could have been conveyed with no length restrictions. In such cases, **multimodal information** must be exploited as much as possible to prioritize the most relevant highlights and generate optimal summaries within the text limitations.
- Another shortcoming of existing approaches is the lack of **fully automated** and **end-to-end** operation. Most works employ manual efforts at one stage or another, and often cannot generate full summaries directly from raw input without human intervention.

In this work, we aim to create an end-to-end and fully automated pipeline which can address the shortcomings listed above.

3 PROPOSED PIPELINE

Our proposed pipeline for generating text summaries from soccer game multimedia is depicted in Figure 1. The pipeline uses 3 types of inputs in 2 (3) modalities:

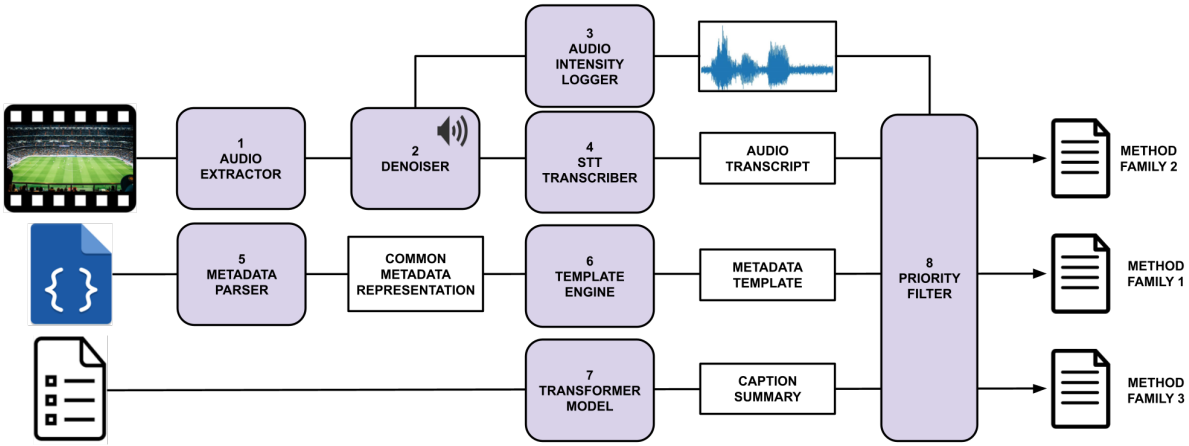

Figure 1: Proposed pipeline for generating text summaries from soccer game multimedia.

- **Game broadcast - audio (and video):** The raw multimedia stream from the broadcast production. It is possible to use only the audio stream if separately available, instead of the complete broadcast including audio and video streams.
- **Game metadata - text:** File(s) with event annotations, indicating certain highlights from the game along with additional information (such as a timestamped goal event and the name of the scoring player, or a free kick and the name of the opposing player causing the awarding of the kick). Such metadata is commonly generated by tagging centres in commercial operation [4], or by research groups [34].
- **Game captions - text:** File(s) including timestamped entries describing the actions happening on the game field. Unlike metadata entries, these are full sentences in natural language form. The scope of such input can be variable, ranging from the basic listing of important game events (such as goals, free kicks, etc., as can be found in the metadata) to more elaborate explanations containing information on the context of the particular game (such as the history of the home and away teams, notes of the status in the particular league or championship, etc.) as well.

3.1 Component Operations

Our proposed pipeline comprises 8 main modules.

3.1.1 Audio extraction. This module is for extracting an audio stream from the video stream. We use `ffmpeg` [53], a popular software framework for transcoding multimedia files including audio and video. Popular video formats include `MP4`, `MOV`, `AVI`, `FLV`, and `MKV`. It handles anything from the oldest and most esoteric formats to the newest and most up-to-date ones. We set the audio bitrate to 128k and the audio sampling frequency to 44100Hz. This particular configuration is used for being the default configuration for Spleeter. The pre-trained audio separation model has been trained with audios with a sampling frequency of 44100Hz.

3.1.2 Background/noise removal. This module is for de-noising the extracted audio. As mentioned above, we use Deezer spleeter [23] for this purpose. It will output separated vocals and accompaniment files. The vocal audio file contains commentary audio with filtered background noises. Note that any other industry standard noise suppression mechanism can be used instead of Spleeter in this module. Depending on the presence of commentary audio in the video, noise suppression can be omitted (e.g., for videos where commentary is not present and the audio intensity during the game is the only aspect of interest).

3.1.3 Audio intensity log generation. This module records the audio intensity, which will later be used to filter the temporally annotated metadata or captions. The time information corresponding to a certain level of audio energy is used to pinpoint the exact timeframe in the video and the extracted audio. A configurable number of audio frames are analyzed to retrieve the intensity level. As rising audio intensity levels generally follow important highlights (aftermath), the selected set of frames mostly contain audio levels after the event has occurred, with additional few frames from before the event also included.

As part of our pipeline, we implement an easy-to-use dashboard for understanding the correlation between audio intensity levels and the events in the game video. Figure 2 presents a screenshot from our audio intensity analysis dashboard, which 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 the filtering step (Section 3.1.8). We provide this tool as open-source software for the community.

3.1.4 Audio transcription using STT. In this module, the noise-free audio commentary is processed by a speech recognition system to convert the audio into text. Alternative third-party tools which can be used for this purpose include Amazon Transcribe [2], Azure Speech to Text [7], Google Speech to Text [6], and IBM Watson Speech to Text [1]. System capabilities such as language support make up an important aspect of generalizability (e.g., the transcription of leagues from non-English speaking countries in Scandinavia, such as the Norwegian Eliteserien or the Swedish Allsvenskan, was not possible using some of the above alternatives), and also present

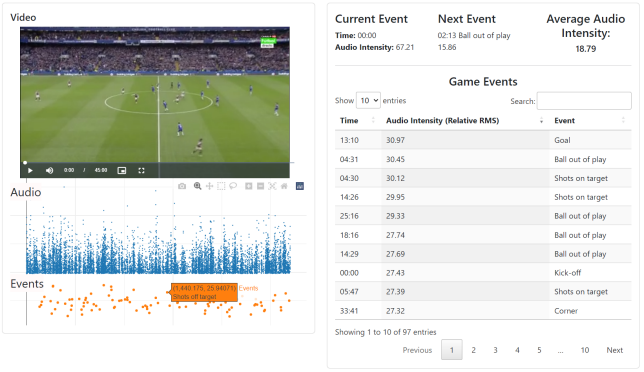

Figure 2: Audio intensity dashboard.

a trade-off between scale and cost. In our current implementation, we use IBM Watson in this module.

3.1.5 Metadata parsing. Different soccer datasets can include game metadata in different formats and use different annotation styles. For uniformity, metadata needs to be parsed and translated into a standard format[2]. This module currently includes support for 3 different metadata input types (in-house dataset, SoccerNet, and K-SportsSum), which can be translated into a common metadata representation.

3.1.6 Template generation. In this module, templates as in Table 1 are applied to the metadata to generate static summaries. The information in the metadata is used to fill the placeholders in the template. The templated are handcrafted for each possible key-set value by analyzing the dataset. Table 1 also presents examples of sentence generation using a naive template, for two different datasets.

3.1.7 Natural language generation with transformers. Language models such as transformers can be used to distil commentary texts into the news. Self-attention, which is the building block of transformers, being a costly operation, limits the total number of words that can be fed to or can be expected as output from the transformers. However, new models such as Longformers [10] replace the standard self-attention operation by local windowed attention with a task-motivated global attention strategy. Such a mechanism for self-attention allows a large number of tokens as the input, making it a good candidate for the summarization of large input texts. Models with 16384 tokens as input and 1024 tokens as output have shown promising performance on multiple downstream tasks, including summarization. The model can be finetuned to generate news from long commentary texts.

3.1.8 Priority filtering using audio intensity. In this optional module, prioritization rules are used to weigh candidate sentences or particular events in the game that would be included in the summary. Audio intensity levels as identified by a previous module (see Section 3.1.3) are utilized. Such a prioritization helps the overall pipeline create a maximally informative summary for a given set of conditions, such as length constraints.

[2]In cases where only game commentaries or captions are available (metadata not available), trained language models can be used to construct metadata. Such a process can help enrich the information available, as well as filter out irrelevant or unimportant content. A fine-tuned GPT-3 [11] model is a good candidate for generating metadata directly from machine-generated captions and human commentary. See Section 4.2

3.2 Alternative Summarization Methods

In this section, we describe alternative end-to-end summarization methods which can be run using our pipeline. Table 2 presents an overview of these methods, where two adopt naive approaches and the third is our proposed approach.

3.2.1 Method 1: summary from game metadata using naive template. We use the game metadata (examples in Listings 1 and 2) along with a naive template as exemplified in Table 1 to generate text summaries. A priority mechanism based on audio intensity can be employed to explicitly filter important events.

3.2.2 Method 2: summary from game audio using naive STT transcription. STT is employed for getting text transcripts of the human commentary from the game audio. The scope of such texts is generally wider as the audio contains the conversation referring to the history of the team or players, the status of the team in the league, etc. A text-based filtering mechanism can be used to remove redundant sentences, and an audio-based filtering mechanism can be used to identify relevant lines of the transcribed text that are deemed important, for inclusion in the summary.

3.2.3 Method 3: summary from game commentary using transformer model. A transformer model is trained so that, for a given set of time-stamped game commentary texts, it predicts the summary of the game. The input and output limits of this method are constrained by the capabilities of the transformer model used.

4 DATASET CURATION

As there are no public datasets with all types of information, such as raw game multimedia, captions, and event metadata as listed in Section 3, available, we curate a number of new/extended datasets[1].

4.1 SoccerNet

SoccerNet [18] consists of a total of 764 hours of 500 different untrimmed broadcast soccer games annotated with three primary event classes: goal, yellow/red card, and player substitution. The event annotations are manually refined to one-second resolution from the coarse data, which is automatically generated from event reports. This dataset focuses on the localization of sparse events within long game videos. SoccerNet-V2 extends the action classes from 3 to 17 in order to support more advanced event spotting from soccer videos. It also emphasizes temporal segmentation of camera shots and retrieval of the replayed actions in the game. The distribution of the number of occurrences of different types of events per game in the SoccerNet dataset is shown in Figure 3, and the metadata format for a sample event is given in Listing 1.

Ground truth generation: As the SoccerNet dataset is not targeting the game summarization use case, there is no ground truth available for this task. We scrapped news, commentary, lineup information, and match statistics for games from multiple leagues from BBC.com. The link for each game on BBC's website was carefully curated and only 278 games were found on BBC's website. A web crawler was used to extract the above-mentioned information from

Dataset	Metadata Format	Interpretation	Sample Sentence
HOST	('free_kick', 'offending_player', 'team')	d[team][value] was awarded a free kick because of d[offending_player][value].	Bodø/Glimt was awarded a free kick because of Erling Haaland.
SoccerNet	('free_kick', 'team')	d[team][value] was awarded a free kick.	Bodø/Glimt was awarded a free kick.
HOST	('red_card', 'player', 'team')	d[player][value] from d[team][value] got a red card.	Sondre Sørli from Bodø/Glimt got a red card.
SoccerNet	('red_card', 'team')	d[player][value] got a red card.	Bodø/Glimt got a red card.
HOST	('goal', 'assist_by', 'scorer', 'shot_type', 'team')	d[scorer][value] scored a goal by d[shot_type][value] shot for d[team][value] with assistance from d[assist_by][value].	Sondre Sørli scored a goal by right-footed shot for Bodø/Glimt with assistance from Japhet Sery.
SoccerNet	('goal', 'team')	d[team][value] scored a goal.	Bodø/Glimt scored a goal.
HOST	('substitution', 'player_in', 'player_out', 'team)	d[player_in][value] replaced d[player_out][value] in d[team][value].	Patrick Berg replaced Sondre Sørli in Bodø/Glimt.
SoccerNet	('substitution', 'team)	d[player_in][value] replaced one of its players.	Bodø/Glimt replaced one of its players.

Table 1: Template for the generation of naive interpretations from metadata.

the web page. The scope of such scrapped news, although wider and containing non-event conversations as well as information irrelevant to that particular game, can be used for training end-to-end systems. Cleaning methods, as well as relevant news selector modules, can be employed to filter the most relevant sentences of interest. We provide this SoccerNet extension as an open dataset.

Listing 1: Metadata for sample event from the Action Spotting Task of SoccerNet-V2 dataset.

```
{
 "gameTime"   : "<half Number> - <time>",
 "label"      : "<action type>",
 "position"   : "<time in ms>",
 "team"       : "<home/away/not applicable>",
 "visibility": "<visible/not shown>"
 }
```

Method	Input	Family	Denoising	Priority Filt.
1.1	M	Naive	✗	✗
1.2	M+A	metadata	✗	✓
1.3	M+A	template	✓	✓
2.1	A	Naive STT	✗	✗
2.2	A		✓	✗
2.3	A		✗	✓
2.4	A		✓	✓
3.1	C	Transformer model	✗	✗
3.2	C+A		✗	✓
3.3	C+A		✓	✓

Table 2: Alternative methods for end-to-end game summarization with/without denoising and audio-based priority filtering (M: metadata, A: game audio, C: captions).

4.2 SportsSum

SportsSum [25] has a total of 5428 soccer game commentaries with corresponding news scrapped from online sources in Chinese.

Translation: We translated all captions and summaries (news articles) in the dataset from Chinese to English using Azure Cognitive Services Translator [8].

Metadata extraction: The original SportsSum dataset includes captions and ground truth summaries for each soccer game, but not event metadata. We exploited the few-shot learning capabilities of GPT-3 to fine-tune the Text-Davinci-001 model with just a few examples, to directly output metadata containing the game events in JSON format. The translated game commentaries of 2278 examples were converted to metadata using the fine-tuned GPT-3 model. These examples have commentary, extracted event metadata in each of the commentary lines, and news, all in English. The remaining 3150 examples have translated commentary and news only. We also performed high-level manual validation. We provide this SportsSum extension as an open dataset.

4.3 K-SportsSum

K-SportsSum [61] dataset has 7854 sports game summaries together with a large-scale knowledge corpus containing information on 523 sports teams and 14K+ sports players in Chinese. A strict manual cleaning process to denoise news articles had been applied to improve both the scale and the quality of the dataset by the original authors.

Translation: We translated the detailed captions and game summaries for the 7854 samples in the dataset from Chinese to English using Google's Cloud Translation. The captions are in plain English sentences with corresponding timestamps as in the original dataset. Since the metadata extraction has not been performed, the temporal event information is not available. Additionally, 523 sports team information items have also been translated into English. We provide this K-SportsSum extension as an open dataset.

For our experiments, all the translated split-sets in the dataset were combined and filtered such that the character length in the corresponding summary (news article) was less than 2500. Such 7839 samples were again split 80%-20% for training and evaluation.

4.4 HOST (In-house)

This dataset currently consists of 15 complete soccer game videos from the Norwegian Eliteserien, accompanied by a list of highlights in the form of event annotations, for each game. The list of highlights includes annotations for events such as cards, substitutions, shots, goals, etc., along with additional timing metadata. The distribution of the number of occurrences of different types of events per

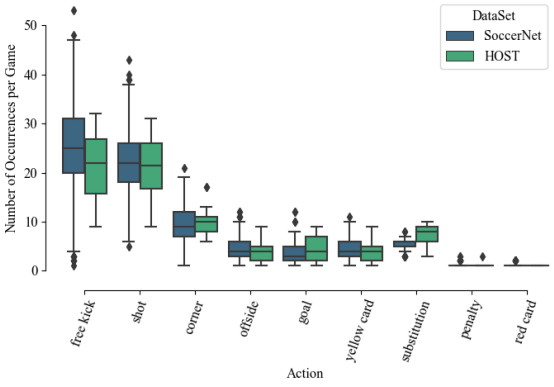

Figure 3: The distribution of the number of occurrences of different types of events per game in SoccerNet and HOST.

game in the HOST dataset is shown in Figure 3, and the metadata format for a card event is given in Listing 2. The dataset is being curated and will be extended further with more samples.

Listing 2: Metadata for card event from the HOST dataset.

```
{'<timestamp>',
    '{"team": {"id": <team-id>,
               "type"  : "team",
               "value" : "<team-name>"},
     "action": "<yellow/red> card",
     "player": {"id": <player-id>,
                "type": "player",
                "value": "<player-name>"}}'}
```

5 EXPERIMENTS

5.1 Metrics

Recall-Oriented Understudy for Gisting Evaluation (ROUGE) is a collection of metrics and software packages to evaluate automated summarization and machine translation software in NLP [16, 21]. It compares the computer-generated summaries to multiple reference summaries written by humans [32]. ROUGE's core concept is to count the amount of overlapping units, such as overlapped n-grams, between candidate (or system) summaries and reference summaries [63]. ROUGE has been proven to be effective in measuring the quality of summaries and correlates well with human judgments [51].

There are multiple ROUGE variants. **ROUGE-1 (R1)** relies on the uni-gram distance between the candidate and reference summaries. **ROUGE-2 (R2)** relies on the bi-gram distance between the candidate and reference summaries. **ROUGE-L (RL)** evaluates the candidate and reference summaries based on their longest common subsequences.

5.2 Preliminary Results

We experiment with the end-to-end summarization methods described in Section 3 using different datasets from Section 4, and the ROUGE metrics as described in Section 5.1. As ground truth, we use the game summaries we have scrapped from BBC.com for SoccerNet, and the English translations we have generated from the game summaries in Chinese for K-SportsSum. For method 3.1, a Longformer model trained for multi-document summarization tasks was finetuned for 10 epochs with the maximum output length set to 1024 over 6 NVIDIA V100 GPUs with a per-device batch size of 6. For the adam optimization with beta(initial decay rates) values of 0.9 and 0.999, and epsilon value of 1e-8, the learning rate was initiated at 5e-5 with linear learning rate schedular. The training took 40 minutes to complete. Table 3 presents our preliminary results in terms of ROUGE-1, ROUGE-2, and ROUGE-L. We present sample outputs in the Supplementary Material document.

Method	Dataset	ROUGE-1	ROUGE-2	ROUGE-L
1.1	SoccerNet	0.08	0.00	0.08
1.2	SoccerNet	0.13	0.01	0.09
2.1	SoccerNet	0.26	0.06	0.10
2.4	SoccerNet	0.29	0.04	0.11
3.1	K-SportsSum	0.52	0.27	0.31

Table 3: Preliminary results for methods 1.1, 1.2, 2.1, 2.4, and 3.1 in terms of ROUGE-1, ROUGE-2, and ROUGE-L metrics.

Our proposed method of using a Longformer (Method 3) demonstrates the potential of using transformers for soccer game video summarization. Although they are limited due to the use of fixed-length input, new architectures are suitable for handling longer-term dependencies in the text in both input and output, making them suitable candidates for text summarization. As seen from the results, these approaches can be superior to naive approaches (Methods 1 and 2), since they are ML based methods able to adapt to the input based on their training and configuration, instead of producing static responses.

6 DISCUSSION

6.1 Comparison with the State of the Art

Our proposed pipeline is the first of its kind in terms of providing an end-to-end automated game summarization functionality requiring no manual intervention in intermediate steps. It is also the first attempt to incorporate audio commentary, metadata, and caption information to provide a comparative analysis. Overall, we see that our approach is competitive with, if not better than, existing work in this domain in terms of the objective ROUGE metric, end-to-end completeness, and automated operation.

The **PASS [59]** framework heavily depends on advanced metadata and needs pre-specified templates. The authors essentially rely on very detailed metadata as input and output a naive template-based summarization (akin to Methods 1.2 and 1.3 presented in Table 2). Our approach directly generates summaries from readily available captions or the ones generated with simple templates without the need for structured game and event details. This work has no ROUGE score as their results are not compared with ground truth summaries. **Huang et al. [25]** have scrapped online sites for events only for their purpose, whereas we undertook web scrapping for commonly used videos in the open SoccerNet dataset with an additional news component for potential use in summarization

tasks. **SportsSum [25]** has scrapped news but has not cleaned them. We provide a more advanced version of SportsSum in English. In **K-SportsSum [61]**, the authors use Selector and Rewriter modules for summarization. They select relevant captions and then rewrite each of them to generate a summary. Our approach takes the input of whole captions at one once and directly outputs summaries without the need for a selector module. This work [61] has attained a maximum ROUGE score of 0.48, which we are able to compete with (see Table 3).

6.2 Potential Applications

The line of research presented in this paper has potential applications in various domains. For instance, **sports companies** who most commonly employ manual pipelines for broadcast soccer production can directly implement solutions such as our proposed pipeline. **Journalists** can benefit from such systems as they can be useful to rapidly generate variable-length game summaries covering important and interesting events, which can be used as news reports on both online platforms fully dedicated to soccer games, as well as sports sections of popular news sites. Systems like Reely [5] and Automated Insights [3] which undertake **automated content creation** can also benefit from automated game summarization, to push updates to social media in real-time for **sport clubs**. Our pipeline can also be directly useful to **online game portals** for whom the game summaries can serve as their video descriptions. As search engines are based on textual information on web content, the text summary of game videos can be very useful in **SEO** and for site indexing such that a search engine can directly point the users from the natural language queries. Such efforts also relate to **information distillation** to address information overload.

6.3 Open Challenges

In this section, we note the limitations and shortcomings of our work, as well as the various open challenges in the domain of soccer game summarization.

Lack of open datasets: As mentioned in Section 4, there are no public soccer datasets which contain the different types of information we have introduced in this work, available in a uniform fashion for the research community. To address this challenge, we have extended a number of existing soccer datasets and started curating our own dataset specifically for the summarization task, but our efforts are far from being adequate on a community scale.

Integration of multiple modalities: The use of multimodal information for the summarization task is still in its infancy. We have used the game audio along with text data for filtering operations based on audio intensity, but more advanced approaches can be investigated.

Heterogeneity in game metadata: Although metadata generation is part of many commercial broadcast pipelines, researchers are still far from having access to such information through commercial open datasets. In cases when game metadata is available, it is mostly sparse and heterogeneous among datasets in terms of the types of events available, event details, and reporting format. Standardization in this respect could help both the research community and commercial stakeholders.

Multilingual operation: Multiple languages in game broadcasts, as well as news articles, can be a problem for summarization systems using out-of-the-box STT tools, as large-scale multilingual support is not yet common. However, STT and translation systems are evolving to handle a wide array of languages, giving hope that future models will have multilingual understanding and generation capabilities accommodating inputs and outputs in multiple languages.

Noise removal: Various types of noise are inherently present in audio streams from sporting events. Prevailing noise removal mechanisms are not suitable for direct use in soccer games. More custom, domain-adapted methods to remove noise effectively from soccer game audio streams would be beneficial for further downstream tasks like STT.

6.4 Future Work

First and foremost, we would like to continue our dataset curation efforts, which also include the analysis, cleanup and validation of the translations of multiple datasets. Generating metadata from existing audio commentary or captions is an integral part of this effort toward richer datasets. We would also like to explore human augmented vs. fully automated methods to clean the news scraped from the internet, which usually has a large scope and might contain information irrelevant to the game. Secondly, we would like to study the effectiveness of using a noise-reduction system before doing STT in more depth, in particular using our audio intensity dashboard introduced in Section 3.1.3. More generally, an ablation study could be performed to quantify the contribution of each change made in the subsequent versions of methods. Thirdly, we would like to investigate the performance of transformer models in more detail, by exploring the efficiency of using a selector module to select relevant commentaries before feeding them to the transformer model to limit input size, as well as mechanisms to explicitly control the output of the text generator module. Last but not least, we are working on establishing a user survey framework to validate our pipeline outputs, as subjective studies are more informative about the end-user experience than objective scores such as ROUGE.

7 CONCLUSION

In this paper, we presented our work in progress on the automatic summarization of soccer games in text format. Through the use of a new end-to-end summarization pipeline, we explored alternative ways for the generation of summaries based on raw game multimedia, as well as readily available game metadata and captions where applicable, utilizing NLP and heuristics. We curated and extended a variety of soccer datasets and provided our preliminary findings from the comparative study of different summarization approaches using various input modalities. We believe our work has contributed to addressing the outstanding issues in multimodal summarization in a sports context. Our open-source software, datasets, and preliminary findings can hopefully be used for future research.

ACKNOWLEDGMENTS

This research was funded by the Norwegian Research Council, project number 327717 (AI-producer).

REFERENCES

[1] 2021. IBM Watson - Speech to Text. https://www.ibm.com/cloud/watson-speech-to-text [Online; accessed 23. Jul. 2022].

[2] 2022. Amazon Transcribe – Speech to Text - AWS. https://aws.amazon.com/transcribe [Online; accessed 23. Jul. 2022].

[3] 2022. Automated Insights. https://automatedinsights.com [Online; accessed 23. Jul. 2022].

[4] 2022. Forzify. https://www.forzasys.com/Forzify.html [Online; accessed 23. Jul. 2022].

[5] 2022. Reely: AI-powered Video Highlights at the Speed of Hype for Sports and esports. https://www.reely.ai [Online; accessed 23. Jul. 2022].

[6] 2022. Speech-to-Text: Automatic Speech Recognition | Google Cloud. https://cloud.google.com/speech-to-text [Online; accessed 23. Jul. 2022].

[7] 2022. Speech to Text – Audio to Text Translation | Microsoft Azure. https://azure.microsoft.com/en-us/services/cognitive-services/speech-to-text/#overview [Online; accessed 23. Jul. 2022].

[8] 2022. Translator – Translation Software as a Service | Microsoft Azure. https://azure.microsoft.com/en-us/services/cognitive-services/translator [Online; accessed 23. Jul. 2022].

[9] 2022. Wyscout. https://wyscout.com [Online; accessed 23. Jul. 2022].

[10] Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The Long-Document Transformer. *arXiv* (April 2020). https://doi.org/10.48550/arXiv.2004.05150 arXiv:2004.05150

[11] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 1877–1901. https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

[12] Joao Carreira and Andrew Zisserman. 2017. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 4724–4733.

[13] Chen-Yu Chen, Jia-Ching Wang, Jhing-Fa Wang, and Yu-Hen Hu. 2008. Motion Entropy Feature and Its Applications to Event-Based Segmentation of Sports Video. *EURASIP Journal on Advances in Signal Processing* 2008 (2008). https://doi.org/10.1155/2008/460913

[14] David L. Chen and Raymond J. Mooney. 2008. Learning to sportscast: a test of grounded language acquisition. In *ICML '08: Proceedings of the 25th international conference on Machine learning*. Association for Computing Machinery, New York, NY, USA, 128–135. https://doi.org/10.1145/1390156.1390173

[15] Freddy Chua and Sitaram Asur. 2013. Automatic Summarization of Events from Social Media. *ICWSM* 7, 1 (2013), 81–90. https://ojs.aaai.org/index.php/ICWSM/article/view/14394

[16] Kavita Ganesan, ChengXiang Zhai, and Jiawei Han. 2010. Opinosis: A Graph Based Approach to Abstractive Summarization of Highly Redundant Opinions. *ACL Anthology* (Aug. 2010), 340–348. https://aclanthology.org/C10-1039

[17] Lorenzo Gatti, Chris van der Lee, and Mariët Theune. 2018. Template-based multilingual football reports generation using Wikidata as a knowledge base. In *Proceedings of the 11th International Conference on Natural Language Generation*. Association for Computational Linguistics (ACL), 183–188. https://research.utwente.nl/en/publications/template-based-multilingual-football-reports-generation-using-wik

[18] Silvio Giancola, Mohieddine Amine, Tarek Dghaily, and Bernard Ghanem. 2018. SoccerNet: A Scalable Dataset for Action Spotting in Soccer Videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. 1711–1721. https://doi.org/10.1109/cvprw.2018.00223

[19] Silvio Giancola, Mohieddine Amine, Tarek Dghaily, and Bernard Ghanem. 2018. SoccerNet: A Scalable Dataset for Action Spotting in Soccer Videos. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 1792–179210. https://doi.org/10.1109/CVPRW.2018.00223

[20] Boqing Gong, Wei-Lun Chao, Kristen Grauman, and Fei Sha. 2014. Diverse sequential subset selection for supervised video summarization. In *NIPS'14: Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*. MIT Press, Cambridge, MA, USA, 2069–2077. https://doi.org/10.5555/2969033.2969058

[21] Virendra Gupta and T. J. Siddiqui. 2012. *Multi-document summarization using sentence clustering*. https://doi.org/10.1109/IHCI.2012.6481826

[22] Rafik Hamza, Khan Muhammad, Zhihan Lv, and Faiza Titouna. 2017. Secure video summarization framework for personalized wireless capsule endoscopy. *Pervasive Mob. Comput.* 41 (Oct. 2017), 436–450. https://doi.org/10.1016/j.pmcj.2017.03.011

[23] Romain Hennequin, Anis Khlif, Felix Voituret, and Manuel Moussallam. 2020. Spleeter: a fast and efficient music source separation tool with pre-trained models. *Journal of Open Source Software* 5, 50 (2020), 2154. https://doi.org/10.21105/joss.02154 Deezer Research.

[24] Liang-Chi Hsieh, Ching-Wei Lee, Tzu-Hsuan Chiu, and Winston Hsu. 2012. Live Semantic Sport Highlight Detection Based on Analyzing Tweets of Twitter. In *2012 IEEE International Conference on Multimedia and Expo*. IEEE, 949–954. https://doi.org/10.1109/ICME.2012.135

[25] Kuan-Hao Huang, Chen Li, and Kai-Wei Chang. 2020. Generating Sports News from Live Commentary: A Chinese Dataset for Sports Game Summarization. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (AACL)*.

[26] Harilaos Koumaras, Georgios Gardikis, George Xilouris, Evangelos Pallis, and Anastasios Kourtis. 2006. Shot boundary detection without threshold parameters. *J. Electronic Imaging* 15 (4 2006), 020503. https://doi.org/10.1117/1.2199878

[27] Mitsumasa Kubo, Ryohei Sasano, Hiroya Takamura, and Manabu Okumura. 2013. Generating Live Sports Updates from Twitter by Finding Good Reporters. In *2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*. Vol. 1. IEEE, 527–534. https://doi.org/10.1109/WI-IAT.2013.74

[28] Leo Leppänen, Myriam Munezero, Mark Granroth-Wilding, and Hannu Toivonen. 2017. Data-Driven News Generation for Automated Journalism. *ACL Anthology* (Sept. 2017), 188–197. https://doi.org/10.18653/v1/W17-3528

[29] Baoxin Li, Hao Pan, and Ibrahim Sezan. 2003. 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).*, Vol. 3. IEEE, III–169.

[30] Ping Li, Yanwen Guo, and Hanqiu Sun. 2011. Multi-keyframe abstraction from videos. *ResearchGate* (Sept. 2011), 2473–2476. https://doi.org/10.1109/ICIP.2011.6116162

[31] Maofu Liu, Qiaosong Qi, Huijun Hu, and Han Ren. 2016. Sports News Generation from Live Webcast Scripts Based on Rules and Templates. In *Natural Language Understanding and Intelligent Applications*. Springer, Cham, Switzerland, 876–884. https://doi.org/10.1007/978-3-319-50496-4_81

[32] Elena Lloret, Laura Plaza, and Ahmet Aker. 2018. The challenging task of summary evaluation: an overview. *Lang. Resources & Evaluation* 52, 1 (March 2018), 101–148. https://doi.org/10.1007/s10579-017-9399-2

[33] Adam Marcus, Michael S. Bernstein, Osama Badar, David R. Karger, Samuel Madden, and Robert C. Miller. 2011. Twitinfo: aggregating and visualizing microblogs for event exploration. In *CHI '11: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 227–236. https://doi.org/10.1145/1978942.1978975

[34] Cise Midoglu, Steven A. Hicks, Vajira Thambawita, Tomas Kupka, and Pål Halvorsen. 2022. MMSys'22 Grand Challenge on AI-based Video Production for Soccer. In *13th ACM Multimedia Systems Conference (MMSys'22)*. ACM. https://doi.org/10.48550/ARXIV.2202.01031

[35] Peter Mindek, Ladislav Čmolík, Ivan Viola, Eduard Gröller, and Stefan Bruckner. 2015. Automatized summarization of multiplayer games. In *SCCG '15: Proceedings of the 31st Spring Conference on Computer Graphics*. Association for Computing Machinery, New York, NY, USA, 73–80. https://doi.org/10.1145/2788539.2788549

[36] Conor Molumby and Joe Whitwell. 2019. General election 2019: Semi-automation makes it a night of 689 stories. https://bbcnewslabs.co.uk/news/2019/salco-ge

[37] Pravin Nagar, Anuj Rathore, C. V. Jawahar, and Chetan Arora. 2021. Generating Personalized Summaries of Day Long Egocentric Videos. *IEEE Trans. Pattern Anal. Mach. Intell.* (Oct. 2021), 1. https://doi.org/10.1109/TPAMI.2021.3118077

[38] Olav Andre Nergård Rongved, Markus Stige, Steven Alexander Hicks, Vajira Lasantha Thambawita, Cise Midoglu, Evi Zouganeli, Dag Johansen, Michael Alexander Riegler, and Pål Halvorsen. 2021. Automated Event Detection and Classification in Soccer: The Potential of Using Multiple Modalities. *Machine Learning and Knowledge Extraction* 3, 4 (2021), 1030–1054. https://doi.org/10.3390/make3040051

[39] Jeffrey Nichols, Jalal Mahmud, and Clemens Drews. 2012. Summarizing sporting events using twitter. In *IUI '12: Proceedings of the 2012 ACM international conference on Intelligent User Interfaces*. Association for Computing Machinery, New York, NY, USA, 189–198. https://doi.org/10.1145/2166966.2166999

[40] Will Oremus. 2014. The First News Report on the L.A. Earthquake Was Written by a Robot. *Slate Magazine* (March 2014). https://slate.com/technology/2014/03/quakebot-los-angeles-times-robot-journalist-writes-article-on-la-earthquake.html

[41] Vassilis Plachouras, Charese Smiley, Hiroko Bretz, Ola Taylor, Jochen L. Leidner, Dezhao Song, and Frank Schilder. 2016. Interacting with Financial Data using Natural Language. In *SIGIR '16: Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*. Association for Computing Machinery, New York, NY, USA, 1121–1124. https://doi.org/10.1145/2911451.2911457

[42] Muhammad Rafiq, Ghazala Rafiq, Rockson Agyeman, Seong-Il Jin, and Gyu Sang Choi. 2020. Scene Classification for Sports Video Summarization Using Transfer Learning. *Sensors* 20 (03 2020), 1702. https://doi.org/10.3390/s20061702

[43] Arnau Raventos, Raul Quijada, Luis Torres, and Francesc Tarres. 2014. Automatic Summarization of Soccer Highlights Using Audio-visual Descriptors. arXiv:1411.6496 [cs.IR]

[44] Ehud Reiter, Somayajulu Sripada, Jim Hunter, Jin Yu, and Ian Davy. 2005. Choosing words in computer-generated weather forecasts. *Artif. Intell.* 167, 1 (Sept. 2005), 137–169. https://doi.org/10.1016/j.artint.2005.06.006

[45] Reede Ren and Joemon M. Jose. 2005. Football Video Segmentation Based on Video Production Strategy. In *Proceedings of ECIR - Advances in Information Retrieval*. 433–446.

[46] Beaux Sharifi, Mark-Anthony Hutton, and Jugal Kalita. 2010. Summarizing microblogs automatically. In *Human language technologies: The 2010 annual conference of the north american chapter of the association for computational linguistics*. 685–688.

[47] Zheng Shou, Dongang Wang, and Shih-Fu Chang. 2016. Temporal Action Localization in Untrimmed Videos via Multi-stage CNNs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 1049–1058. https://doi.org/10.1109/CVPR.2016.119

[48] Karen Simonyan and Andrew Zisserman. 2014. Two-Stream Convolutional Networks for Action Recognition in Videos. In *Proceedings of Advances in Neural Information Processing Systems (NIPS)*. 568–576.

[49] M. U. Sreeja and Binsu C. Kovoor. 2019. Towards genre-specific frameworks for video summarisation: A survey. *J. Visual Commun. Image Represent.* 62 (July 2019), 340–358. https://doi.org/10.1016/j.jvcir.2019.06.004

[50] M. Sun, Ali Farhadi, and S. Seitz. 2014. Ranking Domain-Specific Highlights by Analyzing Edited Videos. *undefined* (2014). https://www.semanticscholar.org/paper/Ranking-Domain-Specific-Highlights-by-Analyzing-Sun-Farhadi/5c7adde982efb24c3786fa2d1f65f40a64e2afbf

[51] Rui Sun, Zhenchao Wang, Yafeng Ren, and Donghong Ji. 2016. Query-biased multi-document abstractive summarization via submodular maximization using event guidance. In *International Conference on Web-Age Information Management*. Springer, 310–322.

[52] Dian Tjondronegoro, Yi-Ping Phoebe Chen, and Binh Pham. 2003. Sports video summarization using highlights and play-breaks. In *Proceedings of ACM SIGMM International Workshop on Multimedia Information Retrieval (MIR)*. 201–208. https://doi.org/10.1145/973264.973296

[53] Suramya Tomar. 2006. Converting video formats with FFmpeg. *Linux Journal* 2006, 146 (2006), 10.

[54] Torrens University Australia. 2020. Why the Sports Industry is Booming in 2020 (and which key players are driving growth). https://www.torrens.edu.au/blog/why-sports-industry-is-booming-in-2020-which-key-players-driving-growth

[55] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning Spatiotemporal Features with 3D Convolutional Networks. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*. 4489–4497. https://doi.org/10.1109/ICCV.2015.510

[56] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. 2018. A Closer Look at Spatiotemporal Convolutions for Action Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 6450–6459. https://doi.org/10.1109/CVPR.2018.00675

[57] J. O. Valand, H. Kadragic, S. A. Hicks, V. Thambawita, C. Midoglu, T. Kupka, D. Johansen, M. A. Riegler, and P. Halvorsen. 2021. Automated Clipping of Soccer Events using Machine Learning. In *2021 IEEE International Symposium on Multimedia (ISM)*. IEEE Computer Society, Los Alamitos, CA, USA, 210–214. https://doi.org/10.1109/ISM52913.2021.00042

[58] Joakim Olav Valand, Haris Kadragic, Steven Alexander Hicks, Vajira Lasantha Thambawita, Cise Midoglu, Tomas Kupka, Dag Johansen, Michael Alexander Riegler, and Pål Halvorsen. 2021. AI-Based Video Clipping of Soccer Events. *Machine Learning and Knowledge Extraction* 3, 4 (2021), 990–1008. https://doi.org/10.3390/make3040049

[59] Chris van der Lee, Emiel Krahmer, and Sander Wubben. 2017. PASS: A Dutch data-to-text system for soccer, targeted towards specific audiences. *ACL Anthology* (Sept. 2017), 95–104. https://doi.org/10.18653/v1/W17-3513

[60] Vimeo Livestream Blog. 2022. Streaming Stats - 47 Must-Know Live Video Streaming Statistics. https://livestream.com/blog/62-must-know-stats-live-video-streaming.

[61] Jiaan Wang, Zhixu Li, Tingyi Zhang, Duo Zheng, Jianfeng Qu, An Liu, Lei Zhao, and Zhigang Chen. 2022. Knowledge Enhanced Sports Game Summarization. In *WSDM '22: Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*. Association for Computing Machinery, New York, NY, USA, 1045–1053. https://doi.org/10.1145/3488560.3498405

[62] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. 2016. Temporal Segment Networks: Towards Good Practices for Deep Action Recognition. In *Proceedings of the European Conference Computer Vision (ECCV)*. 20–36.

[63] Shuai Wang, Xiang Zhao, Bo Li, Bin Ge, and Daquan Tang. 2017. Integrating Extractive and Abstractive Models for Long Text Summarization. In *2017 IEEE International Congress on Big Data (BigData Congress)*. IEEE, 305–312. https://doi.org/10.1109/BigDataCongress.2017.46

[64] Peng Xu, Lexing Xie, Shih-Fu Chang, A. Divakaran, A. Vetro, and Huifang Sun. 2001. Algorithms and system for segmentation and structure analysis in soccer video. In *Proceedings of IEEE International Conference on Multimedia and Expo (ICME)*. 721–724. https://doi.org/10.1109/ICME.2001.1237822

[65] Mary Lynn Young and Alfred Hermida. 2015. From Mr. and Mrs. Outlier To Central Tendencies. *Digital Journalism* 3, 3 (May 2015), 381–397. https://doi.org/10.1080/21670811.2014.976409

[66] Hossam Zawbaa, Nashwa El-Bendary, Aboul Ella Hassanien, and Tai-Hoon Kim. 2012. Event Detection Based Approach for Soccer Video Summarization Using Machine learning. *International Journal of Multimedia and Ubiquitous Engineering (IJMUE)* 7 (1 2012).

[67] Hossam M. Zawbaa, Nashwa El-Bendary, Aboul Ella Hassanien, and Ajith Abraham. 2011. SVM-based soccer video summarization system. In *Proceedings of the World Congress on Nature and Biologically Inspired Computing*. 7–11. https://doi.org/10.1109/NaBIC.2011.6089409

[68] Ruochen Zhang and Carsten Eickhoff. 2021. SOCCER: An Information-Sparse Discourse State Tracking Collection in the Sports Commentary Domain. *ACL Anthology* (June 2021), 4325–4333. https://doi.org/10.18653/v1/2021.naacl-main.342

[69] Bin Zhao and Eric P. Xing. 2014. Quasi Real-Time Summarization for Consumer Videos. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2513–2520. https://doi.org/10.1109/CVPR.2014.322