# AI-Based Sports Highlight Generation for Social Media

Cise Midoglu
SimulaMet, Forzasys
Norway

Saeed Shafiee Sabet
Forzasys
Norway

Mehdi Houshmand Sarkhoosh
OsloMet, Forzasys
Norway

Mohammad Majidi
OsloMet, Forzasys
Norway

Sushant Gautam
SimulaMet, OsloMet
Norway

Håkon Maric Solberg
University of Oslo
Norway

Tomas Kupka
Forzasys
Norway

Pål Halvorsen
SimulaMet, OsloMet, Forzasys
Norway

## ABSTRACT

Social media plays a significant role for sports organizations with millions of active fans, but publishing highlights is often a tedious manual operation. With the development of AI, new tools are available for content generation and personalization to engage audiences. We propose an AI-based multimedia production framework for the automatic publishing of soccer and ice hockey highlights on social media, and disseminate our experiences with developing dedicated pipelines for event detection and classification, player detection and tracking, highlight clipping, cropping, thumbnail generation, game summarization, caption generation, and social media sharing.

## CCS CONCEPTS

• **Information systems** → **Multimedia content creation**; • **Computing methodologies** → **Artificial intelligence**; **Machine learning**; • **Human-centered computing** → *Human computer interaction (HCI)*.

## KEYWORDS

AI, multimedia production, detection, tracking, clipping, cropping, thumbnail, summarization, caption, social media, football, hockey

**ACM Reference Format:**
Cise Midoglu, Saeed Shafiee Sabet, Mehdi Houshmand Sarkhoosh, Mohammad Majidi, Sushant Gautam, Håkon Maric Solberg, Tomas Kupka, and Pål Halvorsen. 2024. AI-Based Sports Highlight Generation for Social Media. In *Mile-High Video Conference (MHV '24), February 11–14, 2024, Denver, CO, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3638036.3640799

## 1 INTRODUCTION AND BACKGROUND

Social media has become a crucial platform for sports organizations, engaging millions of fans worldwide [12, 34, 39, 51, 52, 63, 68]. However, content publishing on these platforms remains to be a largely manual and tedious operation. Recent advancements in AI have introduced new tools for content generation and personalization, aimed at enhancing audience engagement [3, 14, 15, 17, 23, 25, 50, 58, 65]. Despite these developments, there are significant limitations in the current state of the art, particularly in the context of automated sports highlight generation for social media.

*MHV '24, February 11–14, 2024, Denver, CO, USA*
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0493-2/24/02
https://doi.org/10.1145/3638036.3640799

**Event detection:** Detecting events in videos is a complex task, and many different approaches, aimed at a large variety of use-cases, have been proposed in literature. For sports, the traditional manual annotation process, where a group of people annotates video segments as they are aired during a live broadcast, is tedious and expensive. For instance, manually annotating soccer events such as goals and bookings requires a person to watch the game and make bookmarks where relevant actions occur in the video, often at the cost of high latency in publishing. Automation is therefore highly sought after, and several approaches have been proposed over the last years [6, 16, 26, 30, 54, 60, 66]. Action spotting has also been a task in the SoccerNet challenge [8] for several years, with the best teams achiving an average-mAP of around 80% in 2023.[1]

**Player detection and tracking:** Tracking in sports poses a significant challenge due to frequent occlusions and unpredictable movement patterns. Various datasets have been created specifically for multi-object-tracking (MOT) in sports [6, 7], where the top-performing trackers achieve a Higher Order Tracking Accuracy (HOTA) score of around 75% [2, 18, 71].

**Clipping:** One of the most time-consuming and expensive operations in sports multimedia production is the extraction of highlight clips, due to the use of manual trimming, where human operators define the start and end of a clip and trim away the unwanted scenes. The amount of existing work on automatic video clipping is limited. Early works [5, 42, 43, 48, 53, 69] present ideas that can be used for AI-based scene boundary detection and clipping, but the results are limited in terms of accuracy and latency.

**Aspect ratio retargeting (cropping):** Various algorithms exist for adjusting video aspect ratios. Content-adaptive reshaping, such as warping, modifies specific image regions while keeping key areas intact [29, 38]. Segment-based exclusion or cropping focuses on important image parts, removing peripheral visual elements [9, 21, 31, 41]. Seam extraction identifies and removes less critical pixels [27, 61]. Apostolidis and Mezaris [3] highlight the use of cropping for video aspect ratio adaptation, essential for reducing

[1] https://www.soccer-net.org/tasks/action-spotting

semantic distortions. Hybrid techniques merge these strategies, combining cropping with reshaping or seam extraction with exclusion [28, 62]. However, these methods might not be suitable for sports videos, where tracking specific objects such as the players, soccer ball, or hockey puck might be vital. Focusing only on visually salient areas, they may overlook the objects of interest, emphasizing semantic integrity over object presence.

**Thumbnail generation:** Thumbnails that capture the essence of video clips make them more attractive to watch and engage viewers, but are time-consuming to generate. Despite the plethora of literature on creating general-purpose thumbnails based on image quality [49], there are very few works on the automatic generation of thumbnails for sports videos in a content-aware manner.

**Summarization and caption generation:** Due to the dynamic nature of sports, variability between games, the need for accurate event and player tracking information, and contextually relevant content analysis, automated game summarization in sports is a multi-faceted challenge, requiring advancements in various AI and machine learning techniques, along with well-curated datasets [8, 16, 36, 37, 59]. Recent advancements in Large Language Models (LLMs) with a sophisticated understanding of language nuances [33, 70], present new opportunities for experimentation in this field, enabling the summarization of entire sports games effectively by processing extended context windows.

**Social media sharing:** Various studies have focused on the extraction and analysis of data from major social media platforms such as Facebook, Twitter, and Instagram, shedding light on user behaviors, preferences, trends, and the dynamics of online social interactions in general [1, 4, 46]. However, there is a noticeable gap in the study of how automated content sharing affects online engagement and content visibility. Direct publication of sports highlights to social media platforms, facilitated through API integration, can enable an efficient and seamless user experience [35].

## 2 PROPOSED FRAMEWORK

In the following, we present our ongoing work regarding the development of pipelines to support 8 key functions, which we believe to be crucial for the development of a comprehensive end-to-end multimedia production framework for automatically sharing sports highlights on social media.

### 2.1 Event Detection and Classification

In [44, 45], we have presented algorithms to detect and classify soccer events in real-time, using 3D convolutional neural networks. Our results show that there is a trade-off between latency (window search length) and accuracy, which can be exploited. Tested on 3 datasets including SoccerNet [16], our algorithms were able to detect events with high recall and low latency, with up to 93%, 90%, and 86% accuracy for goals, substitutions, and cards, respectively. However, we have received industry feedback deeming this performance to be too low for use in a real setting, as 100% (perfect) detection was required for "official" events[2], before any automated method would be considered for deployment as a commercial tool.

Although sports broadcasts contain both video and audio streams, most detection approaches are unimodal and only consider visual information. In an attempt to improve detection accuracy, we have also investigated the possibility of utilizing both modalities [40]. Our analysis has shown that considering audio features (e.g., log-mel spectrogram of audio signal) may improve performance for some events (e.g., goals where audio intensity is more directly correlated with the event, in the form of audience cheer and commentator speech), while being detrimental for other types of events (e.g., player substitutions where there are no distinct audio markers).

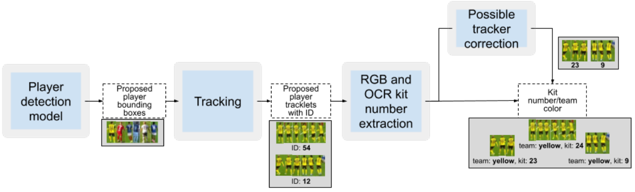


**Figure 1: Player detection and tracking: pipeline.**

**Takeaways and next steps:** We believe that existing approaches for the automatic detection of official events need significant improvement in order to be usable in practice, and that a certain level of manual operation will be required for the foreseeable future, including the verification of automatic detections by humans in a semi-supervised fashion. However, it must be noted that for sports teams and leagues with no resources for live tagging, AI-based automation can still provide considerable benefits. Moreover, highlight clips focusing on non-official events such as tackles, acrobatic shots, dribbles, referee mistakes, and various skill moves are appropriate for social media distribution and quite attractive for fans. Automatic detection approaches are well suited here, as the failure to detect such events is not as catastrophic. We are currently pursuing this direction in our research.

### 2.2 Player Detection and Tracking

The automatic creation of per-player highlight compilations requires the robust tracking of players through long stretches of video. We propose the pipeline presented in Figure 1, which consists of player detection and tracking, followed by team mapping using jersey colors, and optical character recognition (OCR) to extract kit numbers. First, we detect players using the object detection model YOLO [24] fine-tuned on our custom dataset, which yields a bounding box per object in each frame. Next, we associate the bounding boxes throughout frames using Deep-EIoU [18], which employs a re-identification network that produces feature vectors used for similarity measurement. This produces a proposed group of bounding boxes which belong to the same player, referred to as tracklets. Running through these tracklets, we calculate the average RGB value per player and cluster players using DBSCAN++ [22] into two teams based on color. We then run the OCR model PaddleOCR [11] to detect kit numbers. Figure 2 presents a sample result with persistent tracking IDs and predicted kit numbers.

**Takeaways and next steps:** It is possible that the kit number of a player is visible only in a small portion of the frames, yielding no OCR result for some of their bounding boxes. However, if the player tracking is robust, it would suffice to detect the kit number for a few of the boxes in their tracklet. Trying to filter out these "bad crops" where OCR need not be run should therefore be explored. Additionally, running OCR on entire frames instead of only on

[2] Events explicitly described in the Laws of the Game: https://www.theifab.com/.

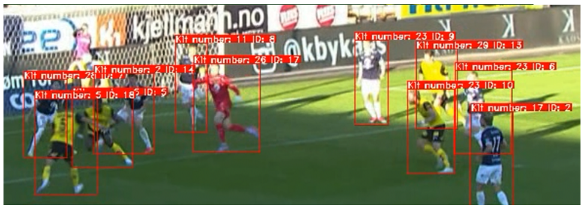


**Figure 2: Player tracking and OCR on sample frame.**

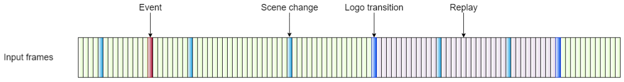


**Figure 3: Conceptualization of sports video flow.**

the bounding boxes within the frames could include players who might be overlooked by the object detection model, improving performance. We are currently working on coupling player tracking with a search/retrieval mechanism, where players of interest can be specified and scenes of a particular player can be clipped.

## 2.3 Highlight Clipping

We conceptualize the general flow of a sports video as a sequence of frames, as depicted in Figure 3. In [55, 56], based on this conceptualization, we have used scene boundary detection, logo detection, and optional cheering removal to automatically generate highlight clips of goal events. Experimenting with different neural network architectures, we found out that the best-performing model for logo detection was a VGG-inspired model for a dataset from the Norwegian Eliteserien league (100% F1-score) and a ResNet model for the SoccerNet dataset (99.7% F1-score). For scene boundary detection, we used a TransNetV2-based model achieving an 87.7% F1-score. Combining these models in an end-to-end event clipping pipeline, our AI-based solution was able to identify scene boundaries, logo transitions, cheering, and replays. Running a subjective user study with 61 participants, we selected 5 sample videos and compared the traditionally used static clipping (e.g., currently in use by Eliteserien) to our automated clipping, with and without cheering and replays. In each question (or "Case"), participants were asked to compare two of these three clipping alternatives in pairwise fashion, applied on the same video. As demonstrated by Figure 4, our pipeline could consistently produce more compelling highlight clips compared to static clipping, and among our methods the shorter alternative was preferred.

Next, we used the same approach for clipping booking events, yellow cards in particular. These are more challenging than goal events, due to the possibility of the foul not appearing on-screen in the original run, but only in the replay, and an additional close up shot of the card being shown by the referee before the replay. Figure 5 presents the user scores from a separate user study, comparing static clipping with our automated clipping, and Figure 6 presents the feedback we received on what viewers prefer to see in a booking highlight clip.

**Takeaways and next steps:** Our results show that AI-based clipping can outperform static clipping with a larger margin for goal events than for booking events. Without loss of generalization, this means that event-specific configurations might be needed for automated clipping pipelines. We are working on customizing these, as well as integrating a *multimodal* approach where game audio is considered alongside the video, and adding support for *interleaved* clipping so that non-event-based compilations (e.g., player-based compilations mentioned in Section 2.2) can also be generated.

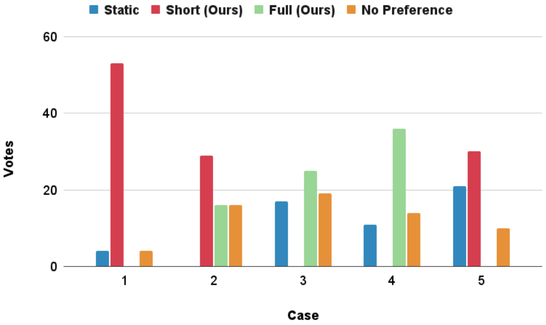


**Figure 4: Clipping of goal events: QoE scores from user study (approaches compared pairwise). Modified from [56].**

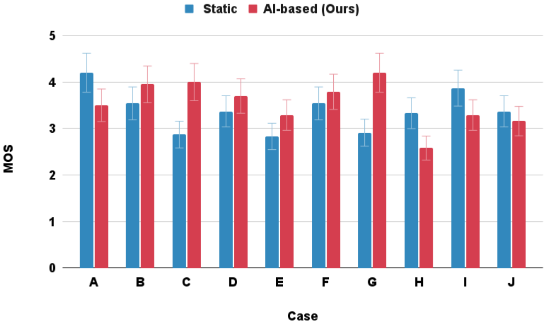


**Figure 5: Clipping of booking events: QoE scores from user study.**

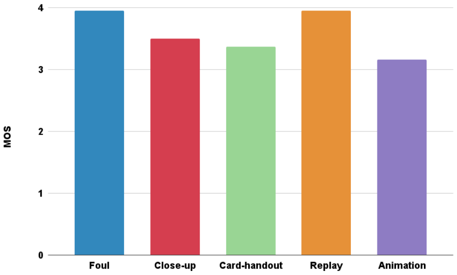


**Figure 6: Clipping of booking events: feedback on viewing preferences from user study.**

## 2.4 Cropping (Aspect Ratio Retargeting)

In [46, 47], we have presented SmartCrop, an automated pipeline for cropping videos to custom aspect ratios supported by various social media platforms. It relies on tracking Points of Interest (POI), with the soccer ball or hockey puck serving as the primary POI. Scenes in the video are identified using the TransNetV2 model, and a YOLOv8-medium object detection model [24, 57], fine tuned on our custom soccer and hockey datasets, is employed to detect the POI. Inaccurate or multiple detections are removed through outlier detection (Z-Score, modified Z-score, or IQR methods). When the POI is visible within a frame, it is used as the center of the cropping window. If the POI is not visible, interpolation is used for soccer (linear, polynomial, ease-in-out, or heuristic methods), while

| Type | Centering | Description | Outl. | Interp./Smooth. |
|---|---|---|---|---|
| 1 | frame-centered | static no padding | ✖ | ✖ |
| 2 | frame-centered | static w. black padding to 16:9 | ✖ | ✖ |
| 3 | ball/puck-centered | use last detected ball/puck position | ✖ | ✖ |
| 4 | ball/puck-centered | w. interpolation/smoothing | ✖ | ✔ |
| 5 | ball/puck-centered | w. outlier detection | ✔ | ✖ |
| 6 | ball/puck-centered | w. interpolation/smoothing and outlier detection | ✔ | ✔ |

**Table 1: Cropping types used in the subjective evaluation (red: soccer, blue: ice hockey).**

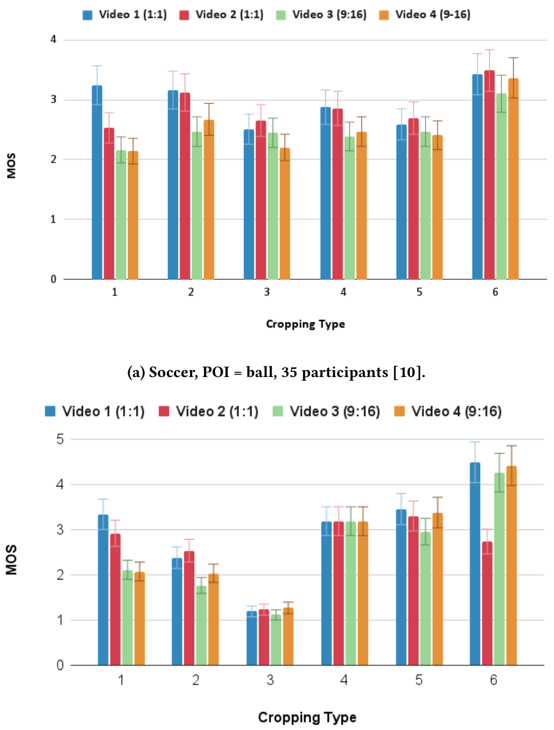
a smoothing module is used to ensure smooth window transitions in ice hockey. Depending on user preferences and POI visibility, the frames are finally cropped either around the POI or the frame's center (static default). We have evaluated SmartCrop for both soccer and ice hockey. In subjective user studies we compared SmartCrop with different types of cropping, as listed in Table 1. We selected two different representative videos in an original aspect ratio of 16:9 from each sport, and cropped them to 1:1 and 9:16 target aspect ratios. For each of these 4 cases, users were asked to rate 6 different cropping alternatives. Figure 7 presents the Mean Opinion Score (MOS) for the overall Quality of Experience (QoE). Overall, SmartCrop (type 6) consistently yields superior results.

**Takeaways and next steps:** The differences between soccer and ice hockey in terms of game pace, shot types, color contrasts, and broadcast video properties have necessitated customized pipelines, which hints that sports-specific approaches may need to be adopted. In the future, we aim to integrate advanced models such as SAM and DINO for semantic segmentation, along with super-resolution techniques. We're also considering customizing the cropping approach based on event type (instead of following the ball or puck in a context-agnostic manner), as well as enhancing player detection, so that off-ball game highlights (interesting actions which do not appear around the ball or puck) can also be cropped with accuracy.

## 2.5 Thumbnail Generation

In [19, 20], we have presented the HOST-ATS automatic thumbnail selection pipeline, which performs logo detection, close-up shot detection, face detection, blur detection and image quality analysis on video frames, in order to rank them as potential thumbnail candidates according to relevance and image quality. To evaluate the output of the pipeline, we performed subjective user studies comparing our AI-based automatic selection with static and manual selection (currently used by Norwegian and Swedish leagues Eliteserien and Allsvenskan). Figure 8 presents the results for static vs. AI-based selection on 13 samples. Overall, we see that the automatic selection is preferred in most cases. Furthermore, we have received feedback on what viewers generally consider as important features of a thumbnail, as depicted in Figure 9.

**Takeaways and next steps:** Automatic thumbnail selection saves human operators a lot of time. Viewers consider high image quality, player faces, and action content as the most important aspects of a thumbnail, followed by close-up shots, cheering, and the absence of logo transitions. These confirm our initial thumbnail selection rules, and give further insights regarding possible additional rules (e.g., detection of action content and cheering context). Based on user feedback, we are working on a postprocessing module to introduce overlays (generating completely new thumbnails from the video frames with added graphics), supporting multimodality.



(a) Soccer, POI = ball, 35 participants [10].

(b) Ice Hockey, POI = puck, 26 participants.

**Figure 7: Cropping user study QoE scores, aspect ratios 1:1 or 9:16.**

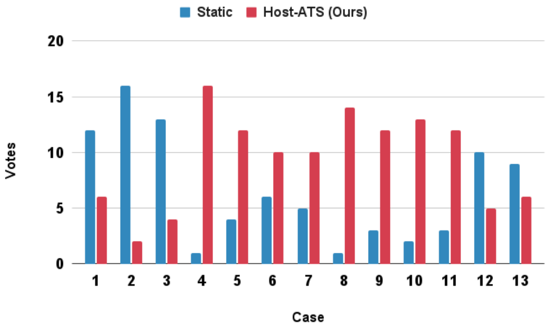


**Figure 8: Thumbnail generation: QoE scores from user study [19].**

We are also looking into techniques for cropping the frames, as well as possibly creating animations (GIFs).

## 2.6 Game Summarization

In [14, 15], we have proposed an approach for crafting comprehensive soccer game summaries by integrating various input modalities, including game audio, structured metadata, and captions. Figure 10 illustrates our automated pipeline, which uses automatic speech recognition to convert game audio into text commentaries, and an adaptive template engine to transform structured event information into natural sentences, ensuring efficient summary generation crucial for broadcasting and journalism. Transformer-based language models distill multimodal input texts into concise summaries. These models efficiently handle large volumes of text, allowing

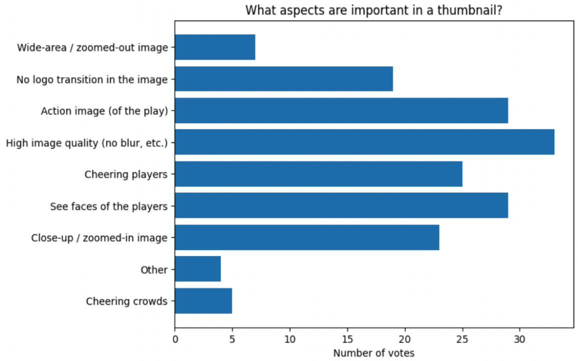


Figure 9: Thumbnail generation: user feedback.

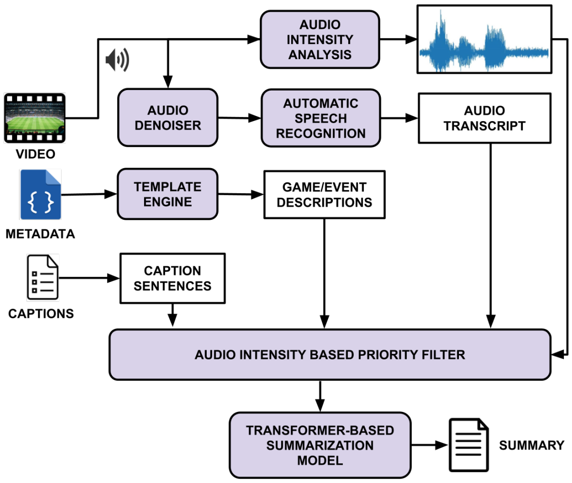


Figure 10: Game summarization: pipeline.

for comprehensive coverage of game events and discussions [15]. A key feature of this system is the analysis of audio intensity to identify exciting game moments, such as goals or penalties. The pipeline also includes a priority filtering module that focuses on significant events, leveraging audio intensity data to highlight the most relevant game moments [14]. We have enriched various soccer datasets with translations and metadata, facilitating the training of data-intensive models, particularly transformers.

**Takeaways and next steps:** To enhance the textual output, we are experimenting with multimodal scene understanding architecture as proposed in [32, 33, 64, 70]. This approach utilizes modality-specific encoders tailored for each type of multimedia content present in the highlights (i.e., video, audio, image, and text). Additionally, incorporating adapter layers that have been trained on an instruction dataset (question-answer pairs), followed by the integration of a conversational Language Model (LLM) can increase model understanding [13]. To further optimize our output, we suggest pertinent hashtags and captions that are generated based on the video content, taking into consideration specific games and current trending topics. This comprehensive strategy ensures improved discoverability of our content among a wider audience, ultimately leading to increased fan engagement [13, 67].

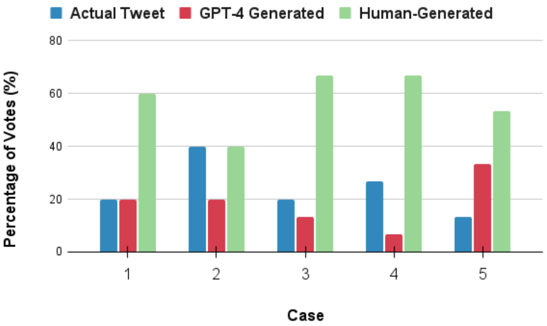


Figure 11: Caption generation: QoE scores from user study.

## 2.7 Caption Generation

Understanding and capturing the dynamic essence of soccer events in captions is crucial for engaging a global audience. In order to evaluate the feasibility of automated caption generation and understand user preferences, we conducted a preliminary user study with 15 participants, where participants were asked to compare, for a particular soccer goal event, (1) an actual tweet from the relevant club about the event, (2) a GPT-4 generated tweet, including only metadata such as the timestamp of the goal, the scorer, and the type of goal, and (3) a human-generated tweet focused on describing the scene in detail. The results showed a clear preference for more descriptive, human-like tweets when discussing soccer goal events (Figure 11).

In response to these insights, we designed an advanced pipeline for generating automated short text summaries of soccer goal events, tailored for social media platforms such as Twitter and Instagram. As shown in Figure 12, the process begins with the selection and preprocessing of a specific event, which involves downloading the corresponding highlight clip in a supported format, such as an HLS playlist (.m3u8) or a standard MP4 file. Extracting the significant frames from a video is a crucial factor, considering the limitations of token inputs in language models, and the necessity of producing comprehensive yet concise content. We apply a shot-type classifier to filter out less relevant frames, such as close-up or medium shots, focusing instead on long and full shots (wider angle views), to allow for more detailed scene descriptions. The selected frames undergo further analysis using specialized models for object detection (with eight distinct classes), pitch segmentation (with two classes), and object tracking (particularly focusing on the goalkeeper). The outputs are compiled into a JSON file, which aggregates the information, reduces redundancy, and logically groups related information pertaining to each frame. This data, along with audio insights from automatic speech recognition (ASR), and processed game metadata, are then fed into GPT-4 using carefully engineered prompts. GPT-4 processes this amalgamated input to produce contextually rich, platform-specific outputs, such as single or multi-threaded tweets and Instagram post captions. Our multifaceted approach ensures that text summaries are not only rich in detail, but also align with the visual and auditory elements of the soccer game, offering a comprehensive and engaging narrative of each goal event.

**Takeaways and next steps:** Enhanced object and player tracking in the pipeline can significantly improve the understanding of

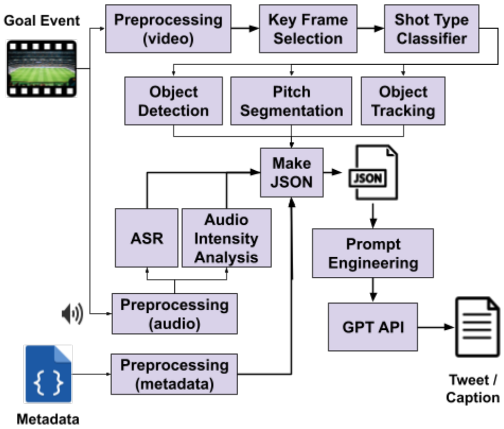


Figure 12: Caption generation: pipeline.

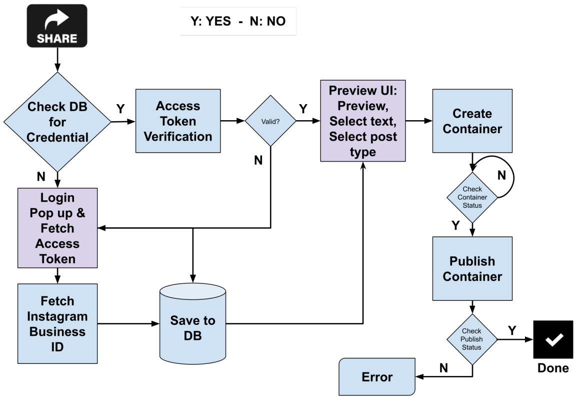


Figure 13: Social media sharing: Instagram workflow.

player movements, ball trajectories, and team formations, offering a more detailed view of the game. By leveraging this data for predictive analysis, the system could anticipate future events such as scoring opportunities or defensive breakdowns, adding an engaging predictive element to the summaries. Developing real-time processing and summarization capabilities for live events would enable the generation of instant updates and comprehensive summaries as the action unfolds. Segmenting the game field into distinct areas such as defensive, midfield, and attacking zones could also provide deeper insights into team strategies and player positions, further enhancing the analytical depth of the summaries.

## 2.8 Social Media Sharing

One of the prevalent challenges is the lack of scalability in the distribution of content across various platforms, and the inefficiency of the download-upload routine. The ideal solution is a system that seamlessly integrates with existing content creation tools, enabling direct and multi-target sharing from a single interface, thus saving time and expanding the reach of content creators. In pursuit of this goal, we present an automated social media sharing pipeline currently supporting TikTok and Instagram (Figure 13 depicts the latter workflow), which negates the need for users to download, manually process, and upload content. The process begins with a robust authentication mechanism allowing secure access to platform features on behalf of users. For communication with social media platform APIs, standardized data formats are utilized, chosen for their lightweight nature and compatibility with common programming languages. Adherence to the API usage policies of TikTok and Instagram is crucial, particularly with respect to rate limits and API quotas. Our pipeline is designed to manage these limits effectively to ensure uninterrupted service and compliance.

**Takeaways and next steps:** As an integrated system, we are currently able to combine automatic highlight clipping, cropping, and social media uploads, with added metadata (caption, hashtags, thumbnails, tags). Our experience has highlighted the importance of comprehensive error handling and monitoring systems for such integrations. These systems are crucial for addressing potential issues such as network problems, invalid responses, or server errors, in order to guarantee reliability and effectiveness. Looking ahead, our focus will be on enhancing the adaptability of our system, in order to keep pace with the constantly evolving landscape of social media platform APIs. This includes integrating automated testing and regular manual reviews to ensure we remain current and responsive to any changes in API features or policies. Next steps involve refining existing processes and exploring the integration of additional social media platforms to expand our reach and capabilities.

# 3 CONCLUSION

In this paper, we propose an AI-based multimedia production framework for automatic sports highlight sharing on social media, based on 8 key functions. We present our existing and ongoing work regarding the implementation of each of these functions in a dedicated pipeline, alongside takeaways and lessons learned. We have tested these components on content from Scandinavian soccer and ice hockey leagues, and ran subjective user studies for evaluation, which show that an AI-based automated approach relying on advanced multimodal scene understanding has a great potential to facilitate social media engagement and increase QoE. Our ultimate goal is to combine these components into a unified framework which allows to generate event/player video highlight compilations (together with attractive thumbnails) or game/event textual summaries, to be shared on selected social media platforms, in a fully automated and end-to-end manner. Future work includes applying the same concept to different sports.

# ACKNOWLEDGMENTS

This research was funded by the Research Council of Norway, project number 346671 (AI-Storyteller).

# REFERENCES

[1] Amelia Acker et al. 2020. Social media data archives in an API-driven world. *Arch. Sci.* 20, 2 (June 2020), 105–123.
[2] Nir Aharon et al. 2022. BoT-SORT: Robust Associations Multi-Pedestrian Tracking. arXiv:2206.14651
[3] Konstantinos Apostolidis et al. 2021. A Fast Smart-Cropping Method and Dataset for Video Retargeting. In *Proc of IEEE ICIP*. 19–22.

[4] Robert Bodle. 2011. Regimes of sharing: Open APIs, interoperability, and Facebook. *Information, Communication & Society* 14, 3 (2011), 320–337.

[5] Chen-Yu Chen et al. 2008. Motion Entropy Feature and Its Applications to Event-Based Segmentation of Sports Video. *EURASIP J. Adv. Signal Process.* 2008, 1 (Dec. 2008), 1–8.

[6] Anthony Cioppa et al. 2020. A Context-Aware Loss Function for Action Spotting in Soccer Videos. In *Proc. of CVPR*.

[7] Yutao Cui et al. 2023. SportsMOT: A Large Multi-Object Tracking Dataset in Multiple Sports Scenes. arXiv:2304.05170

[8] Adrien Deliège et al. 2021. SoccerNet-v2 : A Dataset and Benchmarks for Holistic Understanding of Broadcast Soccer Videos. In *Proc. of IEEE CVPR Workshops*.

[9] Thomas Deselaers et al. 2008. Pan, zoom, scan - Time-coherent, trained automatic video cropping. In *Proc. of IEEE CVPR*. 23–28.

[10] Sayed Mohammad Majidi Dorcheh et al. 2023. SmartCrop: AI-Based cropping of soccer videos. In *Proc. of IEEE ISM*.

[11] Yuning Du et al. 2020. PP-OCR: A Practical Ultra Lightweight OCR System. *CoRR* abs/2009.09941 (2020). arXiv:2009.09941 https://arxiv.org/abs/2009.09941

[12] Terry Eddy et al. 2021. Examining Engagement With Sport Sponsor Activations on Twitter. *International Journal of Sport Communication* 14, 1 (Jan. 2021), 79–108.

[13] Sushant Gautam. 2023. Bridging Multimedia Modalities: Enhanced Multimodal AI Understanding and Intelligent Agents. In *Proc. of ICMI*. 695–699.

[14] Sushant Gautam et al. 2022. Assisting soccer game summarization via audio intensity analysis of game highlights. In *Proc. of IOE Graduate Conference*, Vol. 12. Institute of Engineering, Tribhuvan University, Nepal, 25 – 32.

[15] Sushant Gautam et al. 2022. Soccer Game Summarization using Audio Commentary, Metadata, and Captions. In *Proc. of ACM NarSUM*. 13–22.

[16] Silvio Giancola et al. 2018. SoccerNet: A Scalable Dataset for Action Spotting in Soccer Videos. In *Proc. of IEEE/CVF CVPR Workshops*.

[17] Sujuan Hou et al. 2023. Deep Learning for Logo Detection: A Survey. *ACM Trans. Multimedia Comput. Commun. Appl.* 20, 3 (Oct. 2023), 1–23.

[18] Hsiang-Wei Huang et al. 2023. Iterative Scale-Up ExpansionIoU and Deep Features Association for Multi-Object Tracking in Sports. arXiv:2306.13074

[19] Andreas Husa et al. 2022. HOST-ATS: automatic thumbnail selection with dashboard-controlled ML pipeline and dynamic user survey. In *Proc. of MMSys*. 334–340.

[20] Andreas Husa et al. 2022. Automatic Thumbnail Selection for Soccer Videos Using Machine Learning. In *Proc. of ACM MMSys*. 73–85.

[21] Eakta Jain et al. 2015. Gaze-Driven Video Re-Editing. *ACM Trans. Graphics* 34, 2 (March 2015), 1–12.

[22] Jennifer Jang et al. 2018. DBSCAN++: Towards fast and scalable density clustering. *CoRR* abs/1810.13105 (2018). arXiv:1810.13105 http://arxiv.org/abs/1810.13105

[23] Jun-Gyu Jin et al. 2023. Object-Ratio-Preserving Video Retargeting Framework based on Segmentation and Inpainting. In *Proc. of IEEE/CVF WACVW)*. 03–07.

[24] Glenn Jocher et al. 2023. Ultralytics YOLOv8. GitHub. https://github.com/ultralytics/ultralytics Version 8.0.0.

[25] Kawaljeet Kaur Kapoor et al. 2018. Advances in Social Media Research: Past, Present and Future. *Inf. Syst. Front.* 20, 3 (June 2018), 531–558.

[26] Andrej Karpathy et al. 2014. Large-Scale Video Classification with Convolutional Neural Networks. In *Proc. of CVPR*. 1725–1732.

[27] Harpreet Kaur et al. 2019. Video retargeting through spatio-temporal seam carving using Kalman filter. *IET Image Proc.* 13, 11 (Sept. 2019), 1862–1871.

[28] Stephan Kopf et al. 2011. Algorithms for video retargeting. *Multimed. Tools Appl.* 51, 2 (Jan. 2011), 819–861.

[29] Ho Sub Lee et al. 2019. SmartGrid: Video Retargeting With Spatiotemporal Grid Optimization. *IEEE Access* 7 (Sept. 2019), 127564–127579.

[30] Tianwei Lin et al. 2019. BMN: Boundary-Matching Network for Temporal Action Proposal Generation. In *Proc. of ICCV*.

[31] Feng Liu et al. 2006. Video retargeting: automating pan and scan. In *Proc. of ACM MM*. 241–250.

[32] Haotian Liu et al. 2023. Visual Instruction Tuning. *arXiv* (April 2023). https://doi.org/10.48550/arXiv.2304.08485 arXiv:2304.08485

[33] Muhammad Maaz et al. 2023. Video-ChatGPT: Towards Detailed Video Understanding via Large Vision and Language Models. *arXiv e-prints* (June 2023). arXiv:arXiv:2306.05424

[34] Jeff McCarthy et al. 2022. Social media marketing strategy in English football clubs. *Soccer & Society* (April 2022), 513–528.

[35] Gomathy Nayagam Meenakshi Sundaram. 2014. Building Applications with Social Networking API's. *International Journal of High Performance Computing and Networking* 5 (01 2014), 2070–2075.

[36] Cise Midoglu et al. 2022. MMSys'22 Grand Challenge on AI-based Video Production for Soccer. https://arxiv.org/abs/2202.01031

[37] Hassan Mkhallati et al. 2023. SoccerNet-Caption: Dense Video Captioning for Soccer Broadcasts Commentaries. In *Proc. of the IEEE/CVF CVPR Workshops*. 5074–5085.

[38] Hyunwoo Nam et al. 2020. Jitter-Robust Video Retargeting With Kalman Filter And Attention Saliency Fusion Network. In *Proc. of IEEE ICIP)*. 25–28.

[39] Michael L. Naraine et al. 2019. User engagement from within the Twitter community of professional sport organizations. *Managing Sport and Leisure* (June 2019), 275–293.

[40] Olav Andre Nergård Rongved et al. 2021. Automated Event Detection and Classification in Soccer: The Potential of Using Multiple Modalities. *Machine Learning and Knowledge Extraction* 3, 4 (2021), 1030–1054.

[41] Kranthi Kumar Rachavarapu et al. 2018. Watch to edit: Video retargeting using gaze. *Comput. Graphics Forum* 37, 2 (May 2018), 205–215.

[42] Muhammad Rafiq et al. 2020. Scene Classification for Sports Video Summarization Using Transfer Learning. *Sensors* 20 (03 2020), 1702.

[43] Reede Ren et al. 2005. Football Video Segmentation Based on Video Production Strategy. In *Advances in Information Retrieval*. 433–446.

[44] Olav A Nergård Rongved et al. 2021. Using 3D convolutional neural networks for real-time detection of soccer events. *International Journal of Semantic Computing* 15, 02 (2021), 161–187.

[45] Olav A. Norgård Rongved et al. 2020. Real-Time Detection of Events in Soccer Videos using 3D Convolutional Neural Networks. In *Proc. of IEEE ISM*. 02–04.

[46] Mehdi Houshmand Sarkhoosh et al. 2023. Soccer on social media. *arXiv preprint arXiv:2310.12328* (2023).

[47] Mehdi Houshmand Sarkhoosh et al. 2024. AI-Based cropping of soccer videos for different social media representations. In *Proc. of MMM*.

[48] Karen Simonyan et al. 2014. Two-stream convolutional networks for action recognition in videos. In *Proc. of NIPS*. 568–576.

[49] Yale Song et al. 2016. To Click or Not To Click: Automatic Selection of Beautiful Thumbnails from Videos. arXiv:1609.01388 [cs.MM]

[50] Ivan Sosnovik et al. 2023. Learning to Summarize Videos by Contrasting Clips. *arXiv e-prints* (Jan. 2023). arXiv:arXiv:2301.05213

[51] Sandra Tasevski. 2019. Use of Social Media in Communication Strategies of Premier League Football Clubs. *Sinteza 2019 - International Scientific Conference on Information Technology and Data Related Research* (2019), 244–249.

[52] Galen T. Trail et al. 2017. A longitudinal study of team-fan role identity on self-reported attendance behavior and future intentions. *JAS* 3, 1 (March 2017).

[53] Du Tran et al. 2015. Learning Spatiotemporal Features with 3D Convolutional Networks. In *2015 IEEE International Conference on Computer Vision (ICCV)*.

[54] Du Tran et al. 2018. A Closer Look at Spatiotemporal Convolutions for Action Recognition. In *Proc. of CVPR*. 6450–6459.

[55] Joakim O. Valand et al. 2021. Automated Clipping of Soccer Events using Machine Learning. In *Proc. of IEEE ISM*. 210–214.

[56] Joakim Olav Valand et al. 2021. AI-Based Video Clipping of Soccer Events. *Machine Learning and Knowledge Extraction* 3, 4 (2021), 990–1008.

[57] Chien-Yao Wang et al. 2023. YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors. In *Proc. of the IEEE/CVF CVPR*. 7464–7475.

[58] Hang Wang et al. 2022. PAC-Net: Highlight Your Video via History Preference Modeling. In *Proc. of ECCV*. 614–631.

[59] Jiaan Wang et al. 2022. Knowledge Enhanced Sports Game Summarization. In *Proc. of WSDM*. 1045–1053.

[60] Limin Wang et al. 2016. Temporal Segment Networks: Towards Good Practices for Deep Action Recognition. In *Proc. of ECCV*. 20–36.

[61] Shuai Wang et al. 2020. Multi-Operator Video Retargeting Method Based on Improved Seam Carving. In *Proc. of IEEE ITOEC)*. 12–14.

[62] Yu-Shuen Wang et al. 2010. Motion-based video retargeting with optimized crop-and-warp. *ACM Trans. Graphics* 29, 4 (July 2010), 1–9.

[63] Mathieu Winand et al. 2019. International Sport Federations' Social Media Communication: A Content Analysis of FIFA's Twitter Account. *International Journal of Sport Communication* 12, 2 (June 2019), 209–233.

[64] Shengqiong Wu et al. 2023. NExT-GPT: Any-to-Any Multimodal LLM. *arXiv* (Sept. 2023). https://doi.org/10.48550/arXiv.2309.05519 arXiv:2309.05519

[65] Ziwei Xiong et al. 2023. Dual-Stream Multimodal Learning for Topic-Adaptive Video Highlight Detection. In *Proc. of ICMR*. 272–279.

[66] Huijuan Xu et al. 2017. R-C3D: Region Convolutional 3D Network for Temporal Activity Detection. In *Proc. of ICCV*.

[67] Tiezheng Yu et al. 2023. Generating Hashtags for Short-form Videos with Guided Signals. *ACL Anthology* (July 2023), 9482–9495.

[68] Jin Ho Yun et al. 2020. Drivers of soccer fan loyalty: Australian evidence on the influence of team brand image, fan engagement, satisfaction and enduring involvement. *Asia Pacific Journal of Marketing and Logistics* 33, 3 (July 2020), 755–782.

[69] Hossam M. Zawbaa et al. 2012. Event Detection Based Approach for Soccer Video Summarization Using Machine learning. *International Journal of Multimedia and Ubiquitous Engineering* 7, 2 (May 2012), 63–80.

[70] Hang Zhang et al. 2023. Video-LLaMA: An Instruction-tuned Audio-Visual Language Model for Video Understanding. *arXiv* (June 2023). https://doi.org/10.48550/arXiv.2306.02858 arXiv:2306.02858

[71] Yifu Zhang et al. 2022. ByteTrack: Multi-Object Tracking by Associating Every Detection Box. arXiv:2110.06864