JAIST Repository
https://dspace.jaist.ac.jp/

Title	Visual and Memory-Augmented Soccer Commentary Generation
Author(s)	孫, 浩然
Citation	
Issue Date	2026-03
Type	Thesis or Dissertation
Text version	author
URL	https://hdl.handle.net/10119/20403
Rights	
Description	Supervisor:KERTKEIDKACHORN, Natthawut, 先端科学技術研究科, 修士(情報科学)

JAIST
JAPAN
ADVANCED INSTITUTE OF
SCIENCE AND TECHNOLOGY

Japan Advanced Institute of Science and Technology

Visual and Memory–Augmented Soccer Commentary Generation

2410093 SUN Haoran

Automatic soccer commentary generation aims to close the gap between raw broadcast video and the fluent, tactical, and context-aware narrations produced by human commentators. Recent advances in Large Language Models (LLMs) and Vision-Language Models (VLMs) have shown promise in general video captioning; however, the domain of sports commentary presents unique challenges that remain unaddressed. Existing resources and methods are constrained by two fundamental limitations: (1) available annotations are typically concise and event-focused, offering only a single short sentence per clip (averaging around 24 words in datasets like SoccerNet-Caption). These annotations lack the semantic richness necessary to describe the full visual content of a broadcast segment. Crucially, they are not transcriptions of authentic human commentary but are collected from text-based live reporting platforms designed to log key events rather than narrate visual details. Consequently, such captions often omit contextual cues, visual dynamics, and tactical interpretations central to professional broadcast commentary; and (2) most approaches ignore temporal continuity, treating clips as independent moments and failing to model how past events shape the narrative and tactical interpretation of current plays. To address these issues, we first propose a commentary augmentation pipeline that transforms incomplete event-level annotations into a semantically enriched and structurally standardized dataset tailored to video clips. Building on this supervision, we then develop a generation model that produces informative and context-aware commentary by explicitly leveraging visual features and historical event context.

First, we introduce two manually curated datasets: SN-Short and SN-Long. These datasets are designed to capture the richness of language and contextual continuity essential for expert-level sports narration. SN-Short is a scene-level corpus that enriches the single-sentence annotations commonly used in prior work by integrating additional visual details extracted from corresponding audio transcripts and video context; After filtering out irrelevant metadata, SN-Short contains 2.8k verified video–text pairs covering core soccer events such as crosses, shots, set pieces, goals, and fouls. Building on SN-Short, SN-Long connects events across time by linking each target event with semantically related prior events from the same match half. After careful manual selection and aggregation, SN-Long contains 1,765 target events paired with an average of 2.84 historical context segments (5,006 contextual segments in total), producing longer, context-aware commentaries with substantially greater narrative and tactical depth. Both datasets were

manually verified by experienced annotators to ensure accuracy, fluency, and consistency.

Second, we design a commentary augmentation pipeline to tackle the scarcity of high-quality training data. Specifically, this pipeline transforms incomplete, moment-level captions into richly descriptive and structurally standardized texts. This commentary augmentation pipeline leverage pre-processed visual features together with original brief captions as input, to generate an augmented commentary that appends missing visual semantics while preserving stylistic consistency. By applying this pipeline to an extensive corpus of video segments from SoccerNet-Caption, we produce MatchText, a semantically complete and standardized dataset containing 27,207 video–text pairs from 424 matches. This dataset is intended to provide a high-quality supervision signal for downstream commentary generation.

Third, we propose MatchAware: a memory-augmented, retrieval-enhanced commentary generation model that explicitly models both the current visual event and relevant historical events. MatchAware is structured in three components: (i) an event-level video–language generator that produces an initial description of the current clip by projecting Q-Former visual prefixes into a frozen LLM decoder; (ii) a visual event retriever that maintains a memory bank of past event embeddings and retrieves semantically relevant historical events using a time-aware embedding objective; and (iii) a retrieval-augmented generator that conditions on the current feature, the retrieved historical feature, and the temporal offset embedding to produce a context-aware, analytically richer commentary. The retrieval objective encourages semantically meaningful matches while the temporal embedding ensures the model accounts for recency and temporal distance when integrating historical events.

To identify optimal visual representations and validate the effectiveness of both the commentary augmentation pipeline and MatchAware, we conduct extensive experiments on both of them. We first evaluate different visual feature sources (CLIP, Baidu soccer embeddings, ResNet at multiple frame rates, and C3D) within the commentary augmentation pipeline, finding that multimodal fusion improves standard metrics (BLEU, METEOR, CIDEr) and that Baidu features offered favorable information for MatchText construction in our setup. We then compare MatchAware to strong baselines across the SN-Short and SN-Long benchmarks. The comparison includes Video-LLaMA3 in zero- and few-shot settings, along with the MatchVoice architecture trained with different datasets. Models trained on MatchText consistently outperform those trained on raw SoccerNet-Caption or SN-Short; importantly, adding the retrieval-augmented generator yields substantial gains on SN-Long, showing the value of historical context in

generating coherent, context-aware commentary. Across feature variants, MatchAware trained on MatchText achieves the best scores (e.g., large relative improvements in BLEU, METEOR, and CIDEr over non-retrieval baselines), and also in retrieval recall metrics, indicating effective memory matching.

We further validate MatchAware via human evaluation. On a sampled set of clips from held-out games, experienced soccer annotators rated generated outputs along Accuracy (factual match to visual evidence), Completeness (coverage of the main event and its consequences), and Depth (tactical analysis and narrative coherence). While both MatchAware and baseline models achieve comparable accuracy for core events, MatchAware shows statistically significant improvements in Completeness and Depth — consistent with our goals of producing semantically rich and contextually informed commentaries. Ablations confirm that the retrieval mechanism is the primary driver of these improvements on SN-Long, while the augmented MatchText supervision is critical to improving moment-level generation quality.

Finally, we discuss practical limitations and future directions. MatchAware does not include an explicit player localization/tracking module, which constrains its ability to consistently produce correct player identity mentions; the retrieval mechanism is currently limited within a match half and relies on pre-anchored events, restricting cross-match retrieval or fully autonomous segmentation from raw streams; and, despite being the largest manually verified soccer commentary resources to date, the model's overall capability remains limited by the relatively small scale of the data. Our datasets (SN-Short, SN-Long, and MatchText), commentary augmentation pipeline, and MatchAware model together aim to provide a practical foundation for generating professional, analytically rich soccer commentary that aligns with broadcast visual flow.

In conclusion, this thesis presents a comprehensive method designed to bridge the semantic and contextual gaps in automated soccer commentary. By establishing the SN-Short and SN-Long benchmarks, we provide the first manually verified standards for scene-level detail and event continuity. Furthermore, our Commentary Augmentation Pipeline and the resulting MatchText dataset successfully address the scarcity of high-quality training data, transforming fragmented captions into semantically complete narratives. Finally, the MatchAware model demonstrates that explicitly leveraging historical visual context via memory augmentation is key to producing coherent, context-aware commentary. Together, these contributions demonstrate improvements over existing baselines and validate the effectiveness of integrating historical visual context for more coherent sports narration.