人工知能学会第二種研究会資料
AIチャレンジ研究会
SIG-Challenge-067-02

大規模マルチモーダルモデルを用いたRoboCupサッカー競技のハイライト動画の自動生成

Automatic Generation of Highlight Videos using Large Multimodal Model in RoboCup Soccer

鈴木 丈慈[1]　坪倉 和哉[1]　大橋 玲音[1]　小林 邦和[1*]
Joji Suzuki[1]　Kazuya Tsubokura[1]　Reon Ohashi[1]　Kunikazu Kobayashi[1]

[1] 愛知県立大学 大学院情報科学研究科
[1] Aichi Prefectural University

Abstract: RoboCupには多くのリーグがあり，リーグ毎にルールやタスクが細かく設定されているため，一般の方に試合の様子が分かりにくいという問題がある．そこで本研究では，RoboCupの試合情報を一般の方にわかりやすく伝えるために，ナレーション付きハイライト動画を自動生成する手法を提案する．具体的には，RoboCupの試合映像を大規模マルチモーダルモデルにより分析し，試合状況の理解とハイライト抽出，ナレーション文章を生成する．さらに生成したナレーション文章を音声合成し，ハイライトに付与することで，ナレーション付きハイライト動画の生成を行う．提案手法により生成されたハイライト動画を主観評価した結果，人間の作成したハイライト動画には劣るものの，面白さの観点では一定の魅力があることが示唆された．

1　はじめに

RoboCupは，西暦2050年「サッカーの世界チャンピオンチームに勝てる，自律型ロボットのチームを作る」という夢に向かって人工知能やロボット工学などの研究を推進し，様々な分野の基礎技術として波及させることを目的としたランドマーク・プロジェクトである [1]．毎年世界大会が開催され，日本においても国内大会が実施されており，多くの観客を集めている．

RoboCupには関連研究の推進や技術の波及という側面だけではなく，技術に対する社会の関心を高め将来的な競技人口を増やす側面もある．しかしながら，現状のRoboCupは多くのリーグがあり，リーグ毎にルールやタスクが細かく設定されているため，一般の方にはルールが分かりにくいという問題がある．著者らが以前行ったアンケートでも，RoboCupのイベント会場で必要な情報として，「RoboCupの細かいルール」や「小中学生向けの競技の説明」が必要とされており [2]，一般の方に試合の情報をわかりやすく伝えるシステムが求められる．

これまで，試合の状況を観客に伝える方法としては，RoboCupの実況システムが提案されてきた [3, 4, 5]．実況システムでは試合の全て時間を観戦する必要があるため，試合が多数行われているRoboCupにおいては全ての試合を観戦することは困難である．そこで本研究では，より短時間で試合の様子を伝える方法として試合のハイライト動画に着目し，ナレーション付きハイライト動画の自動生成を試みる．試合動画からイベントを抽出して繋ぎ合わせるだけはなく，ナレーションも付与することで視聴者にわかりやすく情報を伝えることが期待できる．

本研究の貢献は以下の2点である．

- RoboCupにおける試合のハイライト動画を自動生成する手法を提案した
- ナレーション付きハイライト動画自動生成における課題点を明らかにした

*連絡先：愛知県立大学大学院情報科学研究科
〒480-1342 愛知県長久手市茨ケ廻間１５２２－３
E-mail: kobayashi@ist.aichi-pu.ac.jp

2　関連研究

2.1　大規模マルチモーダルモデル

大規模マルチモーダルモデル（Large Multimodal Models）の進展により，動画の理解に関する研究が進められている [6, 7, 8]．例えば，動画に対してキャプションを生成したり [9]，動画を要約する技術が研究されている [10]．本研究では，これらの研究に着想を得て，RoboCupの試合の動画を要約してハイライトとして提供するシステムを提案する．

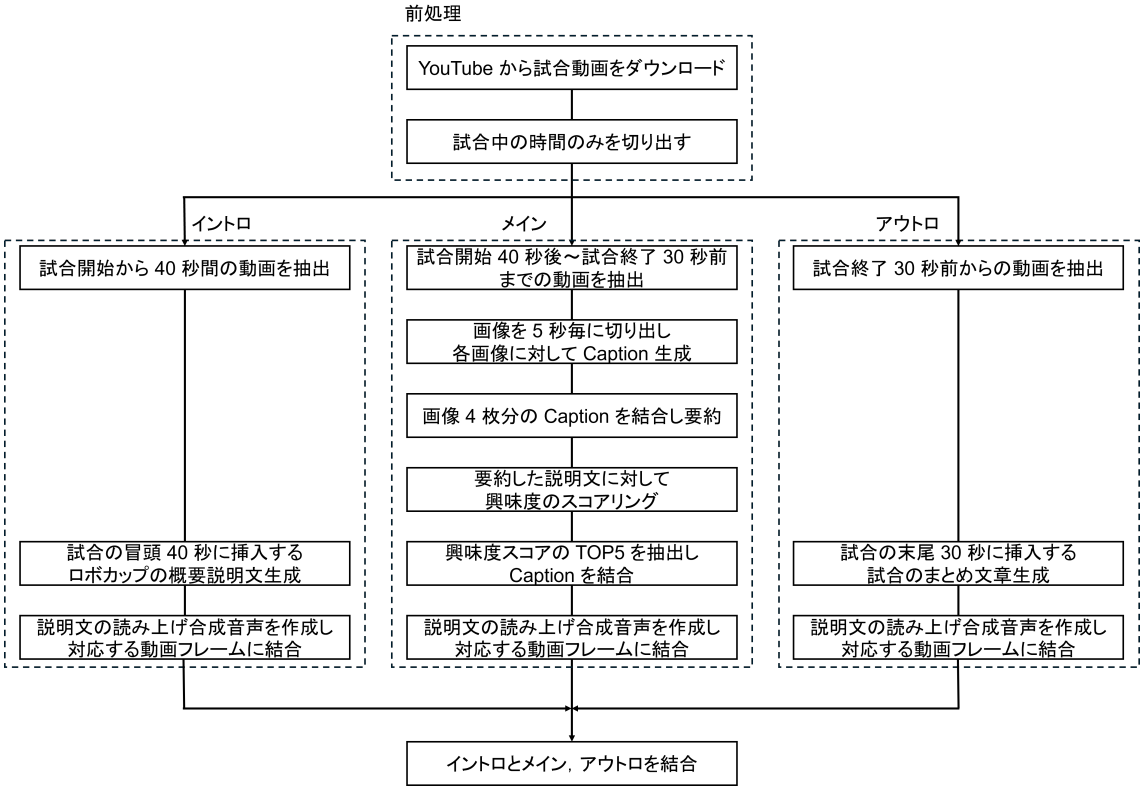

図 1: 提案手法の概要図

2.2 サッカーのハイライト動画生成

これまで，人間同士のサッカーの試合におけるハイライト動画生成に関する研究は行われている [11, 12, 13, 14]．例えば，[13] では，ロゴ検出とシーン境界検出を組み合わせた End-to-End のイベント抽出パイプラインを提案している．また，[14] では，Faster R-CNN を用いたイベント検出手法を提案しており，23 分の試合映像を 4 分 50 秒のハイライト映像へと自動的に圧縮することを可能とした．しかしながら，ナレーション付きでサッカーのハイライト動画を作成する研究は存在せず，またこれらの研究は人間のサッカーを対象としており，RoboCup のハイライト動画の生成に関する研究は行われていない．

3 提案手法

本研究で提案するハイライト動画自動生成システムの実装方法について述べる．提案手法の流れを図 1 に示す．本手法は大きく，前処理部，イントロ部，メイン部，アウトロ部の 4 つに分けられる．RoboCup 標準プラットフォームリーグの試合時間の前半または後半の各 10 分間の試合映像から，イントロ部 40 秒，メイン部 100 秒，アウトロ部約 30 秒の合計 170 秒程度のハイライト動画を作成する．本手法では，大規模マルチモーダルモデルとして OpenAI 社の GPT-4o[15] を，音声合成として gTTS (Google Text-to-Speech)[16] を用いた．各部について，以下で詳述する．

3.1 前処理部

前処理部では，YouTube に公開されている RoboCup 会期中の動画をダウンロードし，試合中の時間のみを切り出す．RoboCup SPL では，試合コートのサイド中央から俯瞰する映像が YouTube チャンネルにて公開されている[1]．チャンネルで公開されている動画は，RoboCup 会期中に行っている数時間から十数時間にわたるライブ配信のアーカイブ動画である．そのため，まず長時間の動画から試合中の動画を抽出する必要がある．ライブ動画のスクリーンショットを図 2 に示す[2]．試合中の映像には，競技中のチーム名や試合の残り時間，試合の点数や状況が含まれている．本研究では，これらの情報を人手で確認して試合が行われている区間を抽出した[3]．

[1] https://www.youtube.com/@RoboCupSPL

[2] 以下の動画のスクリーンショットである．https://www.youtube.com/watch?v=WOA60jGnj5c

[3] OCR により自動で試合時間を抽出することも検討したが，規定された試合ではないテストゲームも抽出されてしまったため，人手で抽出を行った．

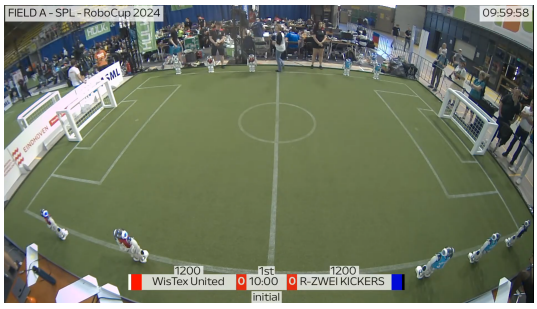

図 2: 試合映像の例

3.2 イントロ部

イントロ部では，試合前半または後半 10 分間の冒頭 40 秒に対して，RoboCup に関する概要を説明する文章を読み上げる．表 1 に示すプロンプトを用いて GPT-4o により説明文章を生成し，gTTS により説明文章の読み上げ音声を生成する．

表 1: イントロ部のプロンプト

あなたは親しみやすいナレーターです。ロボカップサッカーについて、動画の冒頭 30 秒で紹介するナレーションを日本語で作成してください。

3.3 メイン部

メイン部では，試合の前半または後半 10 分間の開始 40 秒後から試合の前半または後半 10 分間が終了する 30 秒前までの区間に対するハイライト動画を作成する．具体的には以下の手順（図 3）で行う．なお，手順の 2 から 5 の処理には GPT-4o を使用した．

1. 試合動画から 5 秒間隔で画像を抽出する．
2. 各画像に対して Caption を生成する．
3. Caption を 4 つずつ結合し，要約する．
4. 要約した文章に対して興味度合いを 1 から 10 段階でアノテーションする．
5. 興味度合いのスコアの最も高い 5 つの文章を選択し，つながりが自然になるように結合する．
6. 結合した説明文章を gTTS により読み上げ音声に変換し，該当する動画区間に付与する．

表 2: メイン部のプロンプト

1 枚画像に対する Caption 生成: あなたはロボットのサッカー競技である「ロボカップ」の試合内容を簡潔に説明するアシスタントです。この画像を 3 文程度で説明してください。画像の中央下部にはチーム名と試合のスコア，残り時間が記されています。これらの情報にも言及してください。{画像}
画像 4 枚分の Caption の要約生成: あなたはロボカップの解説を要約を行うアシスタントです。以下は 20 秒間の動画の 5 秒毎の説明です。これらをまとめて、20 秒間の出来事として簡潔に要約してください。{画像 1 の説明，画像 2 の説明，画像 3 の説明，画像 4 の説明}
興味度合いアノテーション: あなたは映像内容の興味度合いを評価する AI です。次のロボカップの解説文章から、興味度合いをレベルを 1 から 10 で評価してください{解説文章}
説明文章の接続: あなたは親しみやすいナレーターです。次の 5 つの説明文を、つながりのある自然な日本語のナレーション文に書き直してください。{説明文 1，説明文 2，説明文 3，説明文 4，説明文 5}

3.4 アウトロ部

アウトロ部では，試合終了直前 30 秒に対して，試合の結果をまとめる文章を読み上げる．表 3 に示すプロンプトを用いて GPT-4o により説明文章を生成し，gTTS により説明文章の読み上げ音声を生成する．プロンプト中の{画像}は，試合動画の最終フレームを入力とした．

表 3: アウトロ部のプロンプト

あなたはロボットのサッカー競技である「ロボカップ」の解説者です。画像を参考にして試合結果をまとめてください。この試合の結果や印象を 20 秒で語るような日本語のナレーション文を作ってください。{画像}

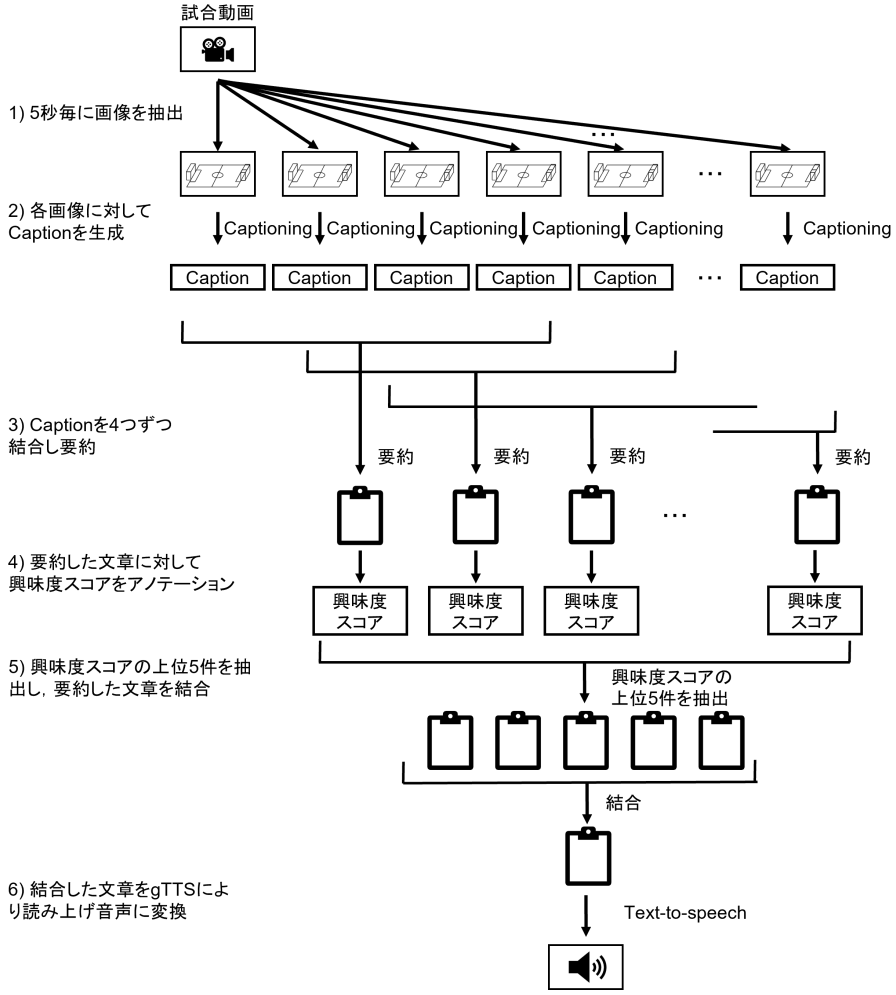

図 3: メイン部の処理の流れ

4　評価と分析

4.1　評価方法

本研究におけるハイライト動画の自動生成システムの有効性を検証するため，大学生と大学院生の実験協力者 12 名を対象に Google Form を用いたアンケート調査を実施した．評価項目は，ハイライト動画の「適切さ」「明瞭さ」「面白さ」「抽出の適切さ」の 4 点である．評価項目の定義を表 4 に示す．評価はいずれも低評価を 1，高評価を 5 とした 5 段階評価とし，人間が作成したハイライト動画を上限の基準とした上で，提案手法によるハイライト動画を評価してもらった．また，自由記述により「人間が作成したハイライト動画と比較して，AI が作成したハイライト動画に不足している点や改善すべき点」を回答してもらった．

4.2　アンケート結果

アンケート結果を表 5 に示す．評価値の 5 が人間が作ったハイライト動画と同等であることを考えると，全ての項目で人間が作成した動画と比較して評価が低いことがわかる．また，評価項目の標準偏差から，明瞭さの評価のばらつきが大きく，実験協力者毎に印象の差があったことが読み取れる．

「人間が作成したハイライト動画と比較して，AI が作成したハイライト動画に不足している点や改善すべき点」に関する自由記述では，主に「動画とナレーションとの不一致」や,「重要なシーン―特にゴールに関連する場面―の抽出不足」が指摘された．具体的には，実験協力者 12 名のうち 7 名が「動画とナレーションの内容が一致していない」と回答し，6 名が「ゴールシーンなどの重要な局面が十分に強調されていない」と述べた．これらの指摘は提案手法のハイライトの自動抽出アルゴリズムの高精度化，および合成音声と動画の時間の対応付けの必要性を示唆している．

表 4: ハイライト動画に対する評価項目と定義

評価項目	質問内容および評価基準（5 段階評価）
適切さ	「AI が作成したハイライトは、適切に試合内容を反映していると思いますか？」
	5: 完全に正確で、試合の内容を忠実に伝えている（人間が作成したものと同等） 1: 全く反映されていない
明瞭さ	「AIが作成したハイライト動画はわかりやすく、明確に理解できると思いますか？」
	5: 非常に明瞭で、誰でも容易に理解できる（人間が作成したものと同等） 1: 非常にわかりづらい
面白さ	「AIが作成したハイライトは興味深く、視聴者の関心を引きつける内容だと思いますか？」
	5: 非常に面白く、強く印象に残る（人間が作成したものと同等） 1: 全く面白くない
抽出の適切さ	「AI が作成したハイライトは元動画から重要な部分を適切に抽出していると思いますか？」
	5: 非常に重要な部分を適切に抽出している（人間が作成したものと同等） 1: 全く重要な部分を抽出していない

4.3 考察

これらの評価結果および実験協力者のコメントを踏まえ，全体として提案手法の生成動画は，人間が作成した動画と比較すると，映像の切り替えやシーン抽出の精度，さらにナレーションのタイミングや内容の明瞭さにおいて課題が残ることが明確となった．これは，5 秒毎の画像の切り出しでは重要なシーンを逃してしまうことや，切り取った画像から状況を正確に判断する能力に不足があるからだと考えられる．一方で,「面白さ」に関しては平均値が 3.33 と，一定の魅力を持つことが示唆されるため，全体の構成自体には一定の評価が得られているともいえる．

以上のことから，提案手法の改善にあたっては，まずナレーションの内容と映像シーンとの連携を強化し，実際の試合の重要な瞬間（特にゴールシーンなど）の抽出アルゴリズムの精度向上が不可欠であることが分かった．また，評価結果のばらつきから，実験協力者間の主観的な違いを踏まえた上で，より客観的な評価基準や，追加的な評価項目を設定することも検討する必要があると考えられる．

表 5: アンケート項目に対する 5 段階評価結果

項目	適切さ	明瞭さ	面白さ	抽出の適切さ
平均値	2.75	2.50	3.33	2.67
標準偏差	0.75	1.17	0.98	0.89

5 むすび

本研究では，RoboCup の試合の様子を短時間で伝える方法として，ナレーション付きハイライト動画を自動生成する手法を提案した．試合映像からハイライトを抽出するだけでなく，状況を説明するナレーションも合成音声として提供することで，ハイライト動画の魅力度向上を目指した．提案手法では，RoboCup 標準プラットフォームリーグの試合動画から GPT-4o を用いて試合状況の理解とハイライト抽出，ナレーション文章の生成を行い，gTTS により合成音声を行う．

実際の試合映像をもとに提案手法と人間が作成したナレーション付きハイライト動画を主観評価した結果,「適切さ」「明瞭さ」「面白さ」「抽出の適切さ」の観点で人間が作成したハイライト動画よりも低い評価となったが,「面白さ」の観点では一定の魅力があることが示唆された．

今後は，映像の切り替えやシーン抽出の精度，ナレーションのタイミングや内容の明瞭さの改善を行う予定である．

参考文献

[1] 野田五十樹，南方英明，小林邦和，杉浦藤虎，武村泰範，秋山英久，岡田浩之，ロボカップ西暦 2050 年を目指して（その１），知能と情報，Vol.29，No.1，pp.2-13，2017.

[2] 坪倉和哉，久保谷空史，早苗昭尚，小林邦和，RoboCup の会場を案内するロボットコンシェルジュの提案，人工知能学会第二種研究会資料，Vol.2022，No.Challenge-059，pp.17-22，2022.

[3] K. Tanaka, H. Nakashima, I. Noda, K. Hasida, I. Frank and H. Matsubara, MIKE: an automatic commentary system for soccer, Proceedings International Conference on Multi Agent Systems, pp. 285-292, 1998.

[4] E. Andre, K. Binsted, K. Tanaka-Ishii, S. Luke, G. Herzog and T. Rist, Three RoboCup Simulation League Commentator Systems, AI Magazine, Vol.21, No.1, pp.57-66, 2000.

[5] 大橋玲音，坪倉和哉，小林邦和，大規模マルチモーダルモデルを用いたロボカップサッカー標準プラットフォームリーグにおける自動実況システム，人工知能学会第二種研究会資料，Vol.2024, No.Challenge-065，pp.1-5，2024.

[6] E. Song, W. Chai, G. Wang, Y. Zhang, H. Zhou, F. Wu, H. Chi, X. Guo, T. Ye, Y. Zhang, Y. Lu, J. N. Hwang and G. Wang, MovieChat: From Dense Token to Sparse Memory for Long Video Understanding, 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024.

[7] B. Lin, Y. Ye, B. Zhu, J. Cui, M. Ning, P. Jin and L. Yuan, Video-LLaVA: Learning United Visual Representation by Alignment Before Projection, the 2024 Conference on Empirical Methods in Natural Language Processing, 2024.

[8] M. Maaz, H. Rasheed, S. Khan and F. Khan, Video-ChatGPT: Towards Detailed Video Understanding via Large Vision and Language Models, the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2024.

[9] H. Wei, Z. Tan, Y. Hu, C. W. Chen and Z. Chen, LongCaptioning: Unlocking the Power of Long Video Caption Generation in Large Multimodal Models, arXiv:2502.15393, 2025.

[10] T. Alaa, A. Mongy, A. Bakr, M. Diab and W. Gomaa, Video Summarization Techniques: A Comprehensive Review, arXiv:2410.04449, 2024.

[11] M. Afzal, J. H. Shah, S. ur Rehman, F. A. Khokhar, M. Yasmin, S. Kadry, Automated soccer event detection and highlight generation for short and long views, Multimedia Tools and Applications, 2024.

[12] A. Narayanan, S. Chuprov, L. Reznik, R. Zatsarenko and D. Korobeinikov, Intelligent Soccer Event Detection and Highlights Generation with Broadcast Cues Integration, International Conference on Machine Learning and Applications (ICMLA), 2024.

[13] J. O. Valand, H. Kadragic, S. A. Hicks, V. L. Thambawita, C. Midoglu, T. Kupka, D. Johansen, M. A. Riegler and P. Halvorsen, AI-Based Video Clipping of Soccer Events, Machine Learning and Knowledge Extraction, 3(4), pp.990-1008, 2021.

[14] N. Darapaneni, P. Kumar, N. Malhotra, V. Sundaramurthy, A. Thakur, S. Chauhan, K. C. Thangeda, A. R. Paduri, Detecting key Soccer match events to create highlights using Computer Vision, arXiv:2204.02573, 2022.

[15] OpenAI, “Hello GPT-4o, `https://openai.com/index/hello-gpt-4o/` (accessed 2025/04/07).

[16] gTTS, `https://pypi.org/project/gTTS/` (accessed 2025/04/07).