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

ＡＩチャレンジ研究会(第 67 回)

Proceedings of the 67th Meeting of Special Interest Group on AI Challenges

CONTENTS

日時:2025年5月4日

場所: 滋賀ダイハツアリーナ

SHIGA DAIHATSU ARENA, May 4th, 2025

一般社団法人　人工知能学会
Japanese Society for Artificial Intelligence

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-067-01

RoboCup Soccer Simulation 2Dを対象とした双方向GRUによる選手とボールの軌道予測

Player and ball trajectory prediction by bidirectional GRU for RoboCup Soccer Simulation 2D

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

[1] 愛知県立大学
[1] Aichi Prefectural University

Abstract: 本研究では，RSS2D を対象として, 選手とボールのゴール時の軌道を予測する双方向 GRU モデルを構築した. 訓練には 2024 年の RSS2D 世界大会に出場した 10 チームの総当たり戦で生成された 90,000 試合のデータから抽出した 237,599 件のゴールのデータを用いた. また, 9 通りのハイパーパラメータの組み合わせを比較することで最も高精度なモデルを選定した. その結果, 最良モデルは双方向 GRU 層が 4 層, 中間層のユニット数が 256 個の構成であることが明らかになった. 最終的に, テストデータに対して, 最良モデルによる予測精度 (終端誤差) は 3.7880 であった.

1　はじめに

RoboCup Soccer Simulation 2D(以下 RSS2D) は, 11 体の仮想的なロボットで構成される 2 チームが対戦するサッカーシミュレーションリーグであり, 複数ロボット間における協調行動の実現を通して, 人工知能や機械学習の研究・開発を推進することが期待されている.

RSS2D おいて, ロボットやボールの移動軌跡を予測することは, 戦術を決定するうえで重要な要素であると考えられる. 特に, 得点に直結するシュートの動きを事前に予測する能力は, 自チームによる攻撃時のみならず, 相手チームの攻撃に対する防御戦略を構築する上でも有用である. これまで, RSS2D においては, パス先の相手の選択予測 [1] やドリブル中の相手チームの移動予測 [2] などの研究が行われてきた.

こうした背景のもと, 選手及びボールの移動軌跡を対象とした予測精度を競う「サッカー軌道予測コンペティション」が開催された[1]. サッカー軌道予測コンペティションは，Robo Cup サッカーシミュレータによって生成された試合データ [3] を用い，選手およびボールの移動軌道の予測精度を競うものである．具体的には，試合開始からゴール決定直前の 3 秒前までのデータを基に，ゴール決定までの選手及びボールの移動経路を予測することが求められる. モデルの性能評価指標として, 終端誤差を採用している.

本稿では，コンペティションにおける予測モデルの構築および評価について報告する．提案手法として Bidirectional Gated Recurrent Unit (双方向 GRU)[4, 5] を用いて選手及びボールの移動経路の予測を行うモデルを構築し, コンペティションのデータセットを用いてモデルの性能評価を行った. その結果, 最良モデルによる終端誤差は 3.7880 となった. なお, 本手法は第 1 回サッカー軌道予測コンペティションにおいて優秀賞を獲得した.

2　データ

本研究に用いたデータは，Robocup 2024 に出場した 10 チーム間の総当たり戦により生成された 2000 試合，すなわち合計 90,000 試合の `.tracking.csv` データである．元データに含まれるカラムを表 1 に示す.

上記データから，ゴール直前 50 フレームのみを抽出してゴールデータを作成した．50 フレームとした理由は, サッカー軌道予測コンペティションにおけるテストデータの最小が 50 フレームであることに基づく. 加えて, 欠損値が含まれるゴールデータは分析に不適と判断して除外した. さらに, 各カラムに対して 0 から 1 の範囲で正規化を実施した．そして, 各チームには整数値を割り当て，20 ビットの One-hot ベクトルへ変換した.

*連絡先: 小林 邦和，愛知県立大学 情報科学部
〒 480-1342 愛知県長久手市茨ケ廻間 1522-3
E-mail: kobayashi@ist.aichi-pu.ac.jp
[1] `https://sites.google.com/view/stp-challenge/`

表 1: カラム一覧

カラム名	説明
#	フレーム番号
cycle	試合時間のカウント
stopped	試合停止中のカウント
playmode	プレイモード
[lr]_name	各サイドのチーム名
[lr]_score	各サイドの得点
[lr]_pen_score	各サイドの延長ペナルティシュートアウトでの得点
b_{x,y,vx,vy}	ボールの位置および速度
[lr][1-11]_t	各選手のプレイヤータイプ ID
[lr][1-11]_{x,y,vx,vy}	各選手の位置および速度
[lr][1-11]_body	各選手の体の向き
[lr][1-11]_neck	各選手の首の向き
[lr][1-11]_vwidth	各選手の視野角
[lr][1-11]_stamina	各選手のスタミナ

また，エージェントの挙動に対して影響が小さいと判断した以下のカラムは分析対象から除外した．

- #
- cycle
- stopped
- playmode
- [lr]_score
- [lr]_pen_score

なお, 元データには左チームが得点した事例と右チームが得点した事例が混在しているため, 本研究では後者については X 座標を反転させ, すべてのデータを左チームがゴールした形式に統一した．この処理によって, データセットの対称性が確保され, 実質的にデータ量を約 2 倍に増やすことが可能となった．

これらの前処理により，最終的に 237,599 件のゴールデータが得られた．このうち 80%（190,079 件）を訓練データ，20%（47,520 件）をテストデータとして用いた．

3 提案手法

本研究では，双方向 GRU を用いて予測モデルの構築を試みた．双方向 GRU は Gated Recurrent Unit(GRU)[4] をもとに, 双方向 Recurrent Neural Network(RNN)[5] の考え方に基づいて, 未来方向と過去方向の双方向に情報のやり取りを行うように変更したモデルである．双方向 GRU は, 交通流量予測 [6, 7] や動画検出 [8] などに用いられている．双方向 GRU 層の概要図を図 1 に示す．ここで, x_i は i フレームでの特徴量を表し, y は出力を表す．入力はゴール直前 50 フレーム中の前 20 フレームのデータとし，後 30 フレームをターゲットデータとして学習を行った．損失関数には予測した 30 フレームすべての特徴量に対する Mean Squared Error (MSE) を用い，最適化手法として Adam[9] を採用した．実装には PyTorch 及び PyTorch-Lightning を用いた．このモデルの入出次元は (バッチサイズ, シークエンス長, 特徴量数) であり, 本実験では入力長は (512, 20, 223), 出力長は (512, 30, 223) である．

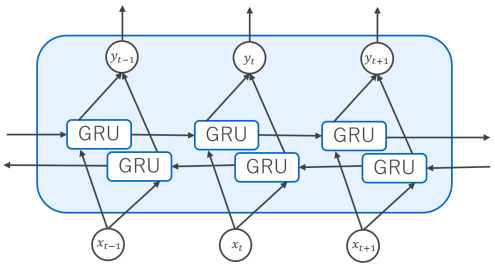

図 1: 双方向 GRU 層の概要図

3.1 ネットワーク構成

提案モデルは，以下の層構成を採用している．提案モデルの概要図を図 2 に示す．

1. **入力層**：各フレームの特徴量をそのまま入力とする．各入力ベクトルは正規化済みの実数値で構成される．
2. **双方向 GRU 層**：双方向 GRU 層を複数層重ねる．
3. **全結合層**：双方向 GRU 層の出力を後続の全結合層に入力し，ターゲットフレームの位置情報を推定する．

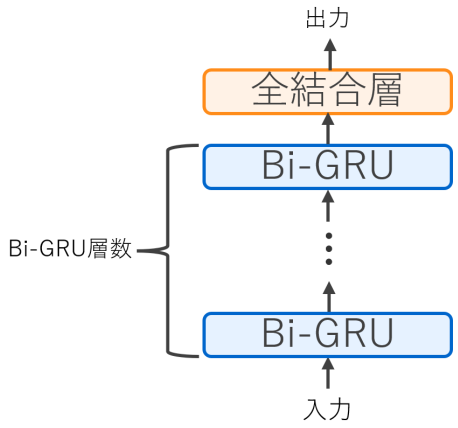

図 2: 提案モデルの概要図

4 実験

実験では，双方向 GRU の層数が 4 層および 8 層，並びに中間層のユニット数が 128 と 256 のモデルを比較検証し，最良の性能を示した構成を採用した．全 200 エポックにわたる訓練を実施し，テストデータにおいて最小の終端誤差を示したエポックのモデルを最終的なモデルとした．終端誤差は，コンペティションの最終的な評価に用いられる評価指標であり，ゴール決定フレームにおけるボールおよび得点側選手の位置におけるユークリッド距離の誤差の平均である．具体的な式を以下に示す．

$$\frac{\sqrt{(x_{\text{acc,ball}} - x_{\text{pred,ball}})^2 + (y_{\text{acc,ball}} - y_{\text{pred,ball}})^2}}{12}$$

$$+\frac{\sum_{i=1}^{11} \sqrt{(x_{\text{acc},i} - x_{\text{pred},i})^2 + (y_{\text{acc},i} - y_{\text{pred},i})^2}}{12}$$

ここで，$x_{\text{acc,ball}}$ および $y_{\text{acc,ball}}$ は実データのボールの x 座標と y 座標を，$x_{\text{pred,ball}}$ および $y_{\text{pred,ball}}$ は予測データのボールの x 座標と y 座標を，$x_{\text{acc},i}$ および $y_{\text{acc},i}$ は実データの i 番目のプレイヤーの x 座標と y 座標を，$x_{\text{pred},i}$ および $y_{\text{pred},i}$ は予測データの i 番目のプレイヤーの x 座標と y 座標をそれぞれ表す．

4.1 ハイパーパラメータ設定

モデルの学習には，以下のハイパーパラメータを設定した．ただし，これらの値は経験則的に求めたものであり，最適なパラメータではない可能性がある．

- **学習率**：Adam オプティマイザを用い，初期学習率は 0.001 とした．
- **バッチサイズ**：512 件ずつのバッチで学習を実施した．
- **エポック数**：200 エポックで訓練を行い，各エポック毎にテストデータで終端誤差を評価した．

4.2 ネットワーク構成

本実験では，モデル構築における各ハイパーパラメータ（双方向 GRU の層数，中間層のユニット数）について，表 2 の設定でグリッドサーチを行い，最適なハイパーパラメータを求めた．

表 2: ネットワーク構成一覧

モデル構成	双方向 GRU の層数	中間層のユニット数
モデル 1	4 層	128 ユニット
モデル 2	4 層	256 ユニット
モデル 3	4 層	512 ユニット
モデル 4	6 層	128 ユニット
モデル 5	6 層	256 ユニット
モデル 6	6 層	512 ユニット
モデル 7	8 層	128 ユニット
モデル 8	8 層	256 ユニット
モデル 9	8 層	512 ユニット

4.3 結果

学習過程における訓練データとテストデータの MSE と終端誤差の推移を図 3, 4, 5, 6 にそれぞれ示す．また，各モデルの最も優れていたエポックの評価結果を表 3 および図 7, 8 にまとめる．実験結果から，4 層・256 ユニット数の構成が最も良好な評価結果を示し，テストデータにおける終端誤差の最小値を記録した．また，双方向 GRU 層は 8 層や 6 層より 4 層のほうが評価が高く，中間層のユニット数は 128 個や 512 個より 256 個のほうが優れていることが判明した．また，モデル 3, 5, 6, 7, 8, 9 は学習過程において MSE 及び終端誤差が大きく上昇する現象ことが確認された．

モデル 3, 5, 6, 7, 8, 9 は学習過程において MSE 及び終端誤差が大きく上がった理由として，モデルのパラメータ数が増えたことによる学習の不安定性の増加が原因と考えられる．また，この最良モデルによって得られた予測図の例を図 4.3, 4.3 に示す．それぞれ，実線と o 印が実際のデータで，x 印と点線が予測データを表す．赤色が左チーム，緑色が右チーム，青色がボールをそれぞれ表す．

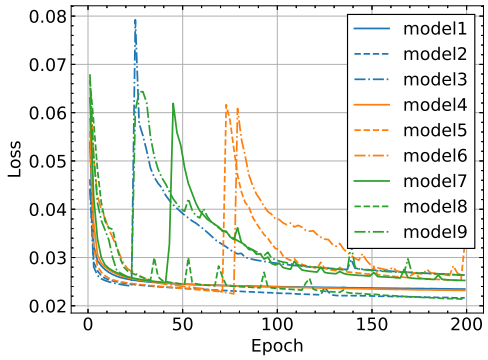

図 3: 訓練データにおける MSE の推移

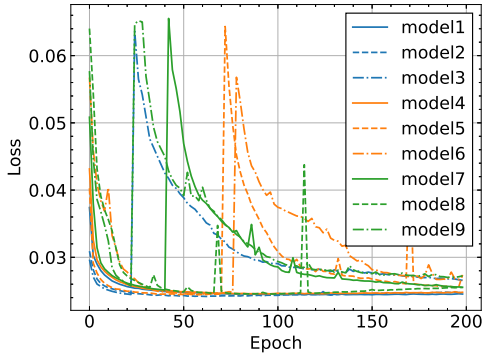

図 4: テストデータにおける MSE の推移

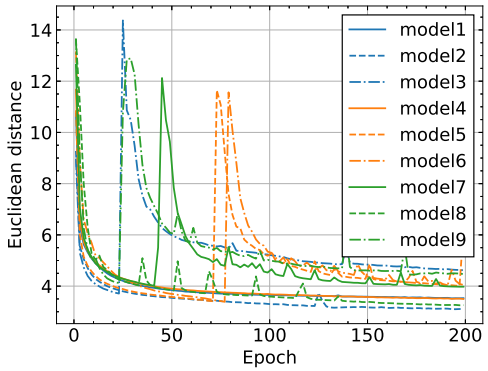

図 5: 訓練データにおける終端誤差の推移

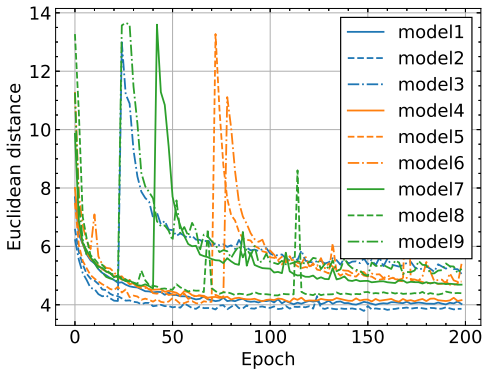

図 6: テストデータにおける終端誤差の推移

表 3: 各モデル構成の評価結果

モデル構成	MSE	終端誤差
モデル 1	0.0244	3.9990
モデル 2	**0.0242**	**3.7880**
モデル 3	0.0245	4.1299
モデル 4	0.0246	4.0942
モデル 5	0.0244	4.0335
モデル 6	0.0245	4.2001
モデル 7	0.0251	4.5604
モデル 8	0.0245	4.3209
モデル 9	0.0258	4.7757

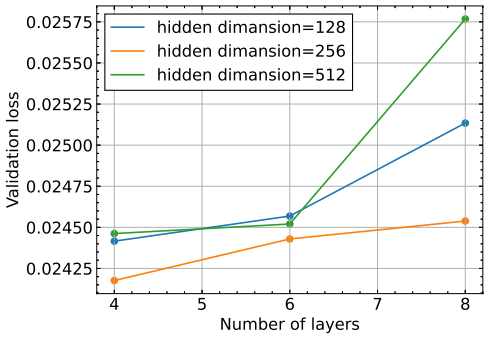

図 7: モデルごとの MSE の比較

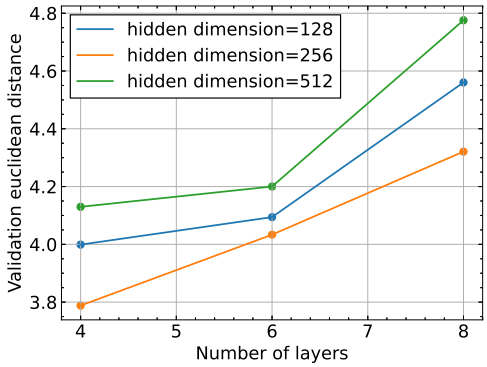

図 8: モデルごとの終端誤差の比較

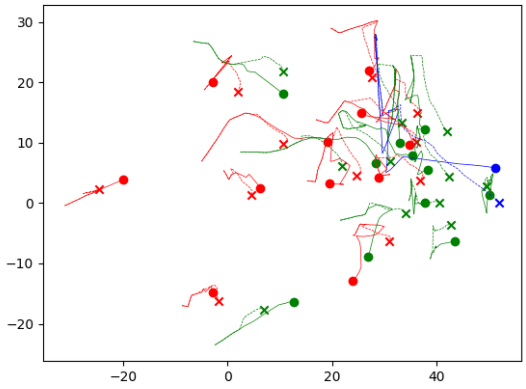

図 9: 予測図 (例 1)

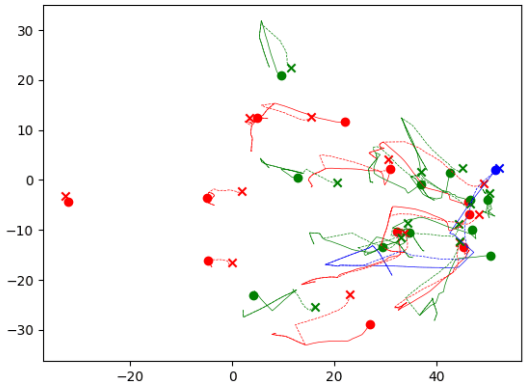

図 10: 予測図 (例 2)

5 まとめ

本研究では，双方向 GRU を用いた RoboCup サッカー軌道予測モデルを構築し，複数のハイパーパラメータを比較することで最も高精度なモデルを選定した．その結果, 双方向 GRU が RSS2D の選手とボールの軌道予測に適していることと, 最適な双方向 GRU のパラメータは総数が 4 層, 中間層ユニット数が 256 であることが明らかになった．

今後の課題として以下の点があげられる．

- **イベント情報の活用**: RoboCup の試合データのうち, .event.csv にはパスやキックといったイベント情報が含まれるが，本研究ではこれらの情報は活用していない．こうしたイベント情報の組み込みを行うことで, より精度を高めることが期待できる．
- **その他のモデルとの比較**: 今回は時系列データが扱え, かつ比較的計算コストの低い双方向 GRU を用いたが, LSTM[10] や Transformer[11] といった他のモデルと比較する．

今後はこれらの手法の検討を進めることで，さらなる予測精度の向上を目指す．

参考文献

[1] 天野巧巳, 内種岳詞, 岩田員典, 伊藤暢浩．Rss2d における期待ポゼッション値の有効性に関する一考察．人工知能学会第二種研究会資料, Vol. 2022, No. SAI-045, p. 01, 2023.

[2] Dmitrij A Petrunenko and Sergej A Belyaev. Prediction of the opponents actions in soccer simulation based on location of players. In *2024 XXVII International Conference on Soft Computing and Measurements (SCM)*, pp. 299–303. IEEE, 2024.

[3] Hidehisa Akiyama and Tomoharu Nakashima. Soccer gameplay data generation: Toward integrating computer simulations and human sports analysis, 2024.

[4] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

[5] Mike Schuster and Kuldip K Paliwal. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, Vol. 45, No. 11, pp. 2673–2681, 1997.

[6] S. Wang, C. Shao, J. Zhang, Y. Zheng, and M. Meng. Traffic flow prediction using bidirectional gated recurrent unit method. *Urban Informatics*, Vol. 1, No. 1, p. 16, 2022. Epub 2022 Dec 1.

[7] Song Liu, Wenting Lin, Yue Wang, Dennis Z. Yu, Yong Peng, and Xianting Ma. Convolutional neural network-based bidirectional gated recurrent unit–additive attention mechanism hybrid deep neural networks for short-term traffic flow prediction. *Sustainability*, Vol. 16, No. 5, 2024.

[8] Abdarahmane Traoré and Moulay A. Akhloufi. *2D Bidirectional Gated Recurrent Unit Convolutional Neural Networks for End-to-End Violence Detection in Videos*, p. 152–160. Springer International Publishing, 2020.

[9] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[10] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, Vol. 9, No. 8, pp. 1735–1780, 1997.

[11] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, Vol. 30, , 2017.

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．

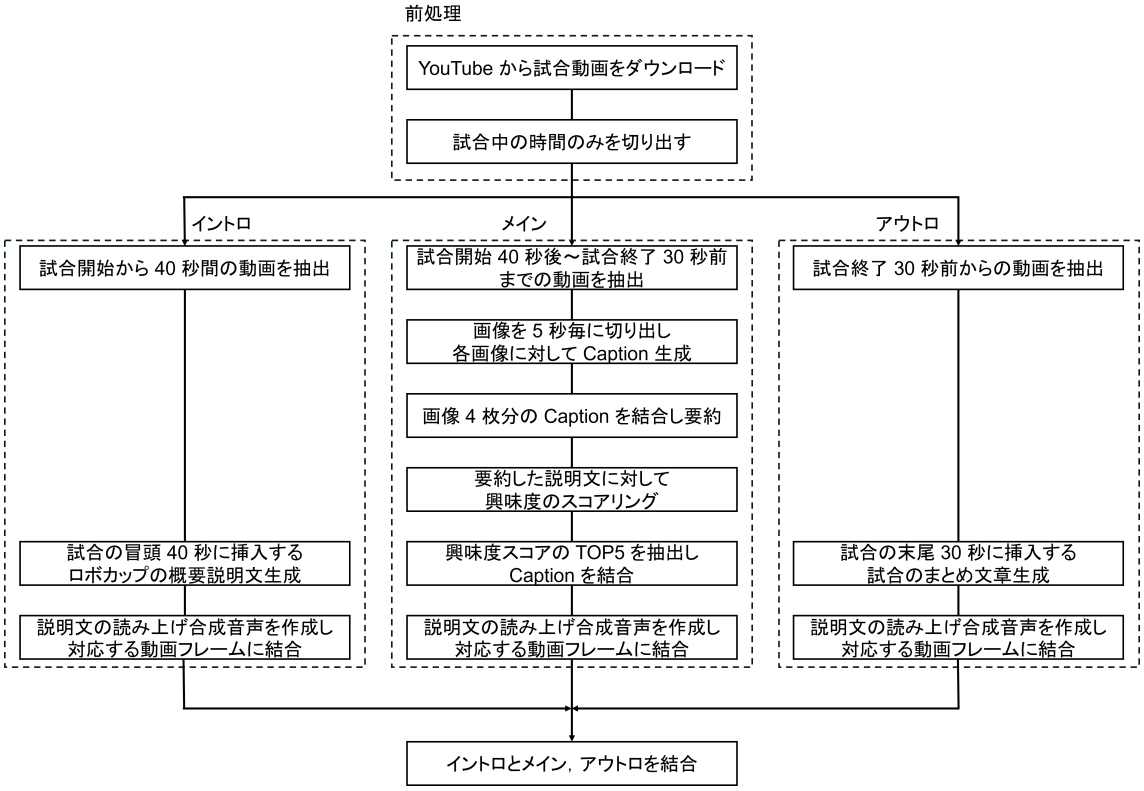

図 1: 提案手法の概要図

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．

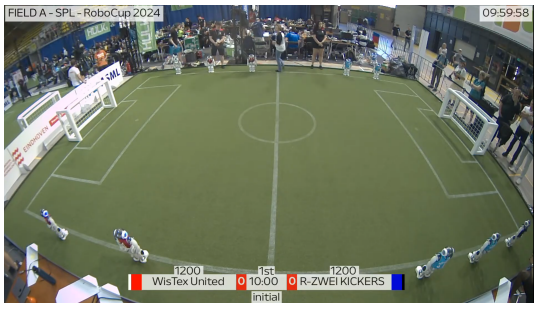

図 2: 試合映像の例

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}

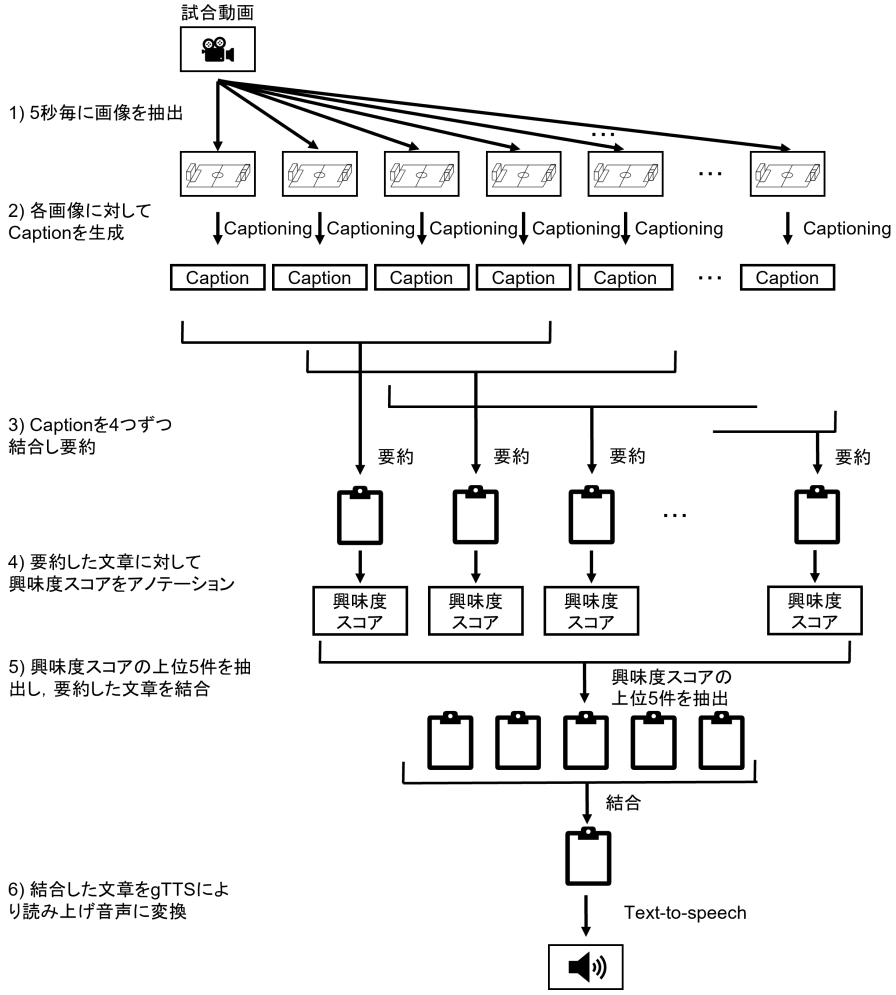

図 3: メイン部の処理の流れ

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).

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-067-03

大規模言語モデルを用いたアーム型ロボットの動作変更に影響を与えるプロンプトの要素に関する一考察

About Elements Which Affect Motion Change of a Robotic Arm Using Large Language Model

安田 尚平[1]　中嶋 洸介[1]　ワルター ニクラス[2]　植村 渉[1]*
Shohei Yasuda[1], Kosuke Nakajima[1], Niclas Walter[2] and Wataru Uemura[1]

[1] 龍谷大学 [1] Ryukoku University [2] ロイファナ大学 [2] Leuphana University

Abstract: 工場や物流倉庫において，組立作業やパレタイズ作業をアーム型ロボットで行うことは一般的である．アーム型ロボットの動作の教示は，専門的な訓練を受けた技術者が行うため，技術者の育成に時間とコストがかかる．アーム型ロボットの動作教示作業にかかるコストの問題を解決するために，大規模言語モデル（LLM）を用いて既存の動作を類似する動作に変更することを考える．その際，動作変更の指示を行うプロンプトを与えても望む動作が得られないどころか，動作変更がなされないことがある．本研究では，その原因を調査するために，プロンプトに含む情報を変えて動作変更をする実験を行い，その結果から動作変更する場合としない場合について考察する．

1　はじめに

工場や物流倉庫における作業では自動化が進み，アーム型ロボットを主としたロボットが活躍している．アーム型ロボットで望む動作を実現するには，動作の教示作業が必要となる．動作の教示作業は専門的な訓練を受けた技術者が行うが，その人材育成には時間がコストがかかる．ロボットやその周辺機器を 1 つのシステムとして統合するためにかかるコストは，ロボット自体のコストも含めた全体のコストの半分以上を占める [1]．ドイツ政府がインダストリアル 4.0 を発表して以降，工場の生産体制は大量生産から多品種少量生産へと移り変わっている．従来のロット生産方式の生産ラインで多品種少量生産体制に対応するには生産ラインの頻繁な段取替えが必要となるため，アーム型ロボットの動作変更にかかるコストを削減することは重要な課題である．

この課題を解決するために，大規模言語モデル（LLM）を用いてアーム型ロボットの既存の動作を類似する動作に変更することを考える．その際，動作変更の指示を行うプロンプトを与えても望む動作が得られないどころか，動作変更がなされないことがある．本研究では，その原因を調査するために，プロンプトに含む情報を変えて動作変更をする実験を行い，その結果から動作変更する場合としない場合について考察する．

2 章では，関連知識として，大規模言語モデルとそれを用いたロボットの制御について説明する．3 章では，LLM を用いてアーム型ロボットの動作を類似する動作に変更する方法を説明する．4 章では，ワークの把持動作を変更する実験を行い，プロンプトに含む情報と動作変更する場合としない場合の違いについて考察する．5 章で本研究をまとめる．

*連絡先：龍谷大学先端理工学部電子情報通信課程
〒 520-2194 滋賀県大津市瀬田大江町横谷 1-5
E-mail: wataru@rins.ryukoku.ac.jp

2　関連知識

2.1　大規模言語モデル

大規模言語モデル（LLM）は大量の自然言語のデータによって学習した確率モデルであり，アテンション機構に基づいた Transformer[2] というモデルを使用するのが主流である．LLM は自然言語で書いたプロンプトを入力するとそれに応じて回答を生成することができ，その応用例として，チャットボット，翻訳，文章の添削，プログラミングへの使用が挙げられる．

2.2　大規模言語モデルを用いたロボット制御

LLM を用いてロボットを制御する研究は注目されている．アーム型ロボットの制御プログラムを生成するには，周辺環境の情報を知覚する重要である．しかしながら，情報が複雑であると文章が長くなり，文章で説明するのは困難である．この問題を解決するために，3 章では，LLM にアーム型ロボットの既存の動作を与え，それに類似する動作に変更する方法を提案する．

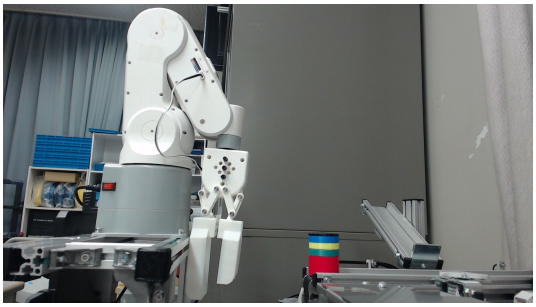

図 1: 実験環境

3　類似する動作への変更

LLM にアーム型ロボットの既存の動作を与え類似する動作に変更する方法を提案する．本手法では，使用可能なコマンドを事前に用意し，それを用いて作成したソースコードがあることを前提とする．そして，LLM にアーム型ロボットを制御する既存のコマンド系列，その動作の説明と欲しい動作の説明を含むプロンプトを入力し，類似する動作を生成する．この方法でアーム型ロボットの動作変更を試みても，望む動作が得られないことや動作変更すらしないことを確認している．そこで，本研究では，実験によって動作変更がある場合とない場合のプロンプトについて調べる．

4　ワーク把持動作の変更実験

4.1　実験方法

動作変更がある場合とない場合のプロンプトについて調査するために，1 図の通りワークの把持タスクを行う環境を用意し，実験を行う．実験タスクとして，自律移動型ロボットの競技大会である RoboCup Logistics League においてロボットが行うワークの把持を扱う．高さ 35mm の円柱状のワークを把持する動作を作成し，高さ 55mm の円柱状のワークを把持する動作に変更するように LLM に指示を出す．本実験では，アーム型ロボットとして，Elephant Robotics 社製のパレタイズロボットの myPalletizer 260Pi，LLM として OpenAI 社の gpt-4o を使用する．

実験手順として，まず，把持と把持戦略の情報，アーム型ロボット本体の情報，グリッパとワークの位置関係の情報の 3 つの情報を含む 8 つのパターンのプロンプトを用意する（表 1 参照）．次に，それらを gpt-4o に入力し，各プロンプトに対して 10 回ずつ出力結果を得る．

表 1: プロンプトのパターンと文字数と動作変更があった回数

把持と把持戦略の情報	アーム型ロボット本体の情報	グリッパとワークの位置関係の情報	文字数 [回]	動作変更があった回数 [回]
無	無	無	209	8
無	無	有	393	10
無	有	無	1135	4
無	有	有	1285	4
有	無	無	425	9
有	無	有	575	10
有	有	無	1317	2
有	有	有	1467	1

4.2　実験結果

4.1 節の実験を行った結果，高さ 55mm のワークを把持する動作に変更できた試行はなかった．動作変更があった回数は表 1 の通りである．

4.3　考察

表 1 からアーム型ロボット本体の情報がない場合はある場合に比べて動作変更があった試行が 3 倍以上多いことがわかる．プロンプトの文字数を見ると，アーム型ロボット本体の情報の文字数は他の情報に比べて 2 倍以上多いことがわかる．このことから，アーム型ロボット本体の情報もしくはその情報を説明するための文字数の多さがアーム型ロボットの動作変更の有無に影響を与えていると考えられる．

5　まとめと今後の課題

本研究では，アーム型ロボットの動作変更にかかるコストの問題を解決するために，LLM を用いてアーム型ロボットの既存の動作を類似する動作に変更する方法を提案した．その際，動作変更がなされないことがあり，その原因を確かめるために，プロンプトに含む情報による動作変更の有無を調べる実験をした．その結果，アーム型ロボット本体の情報もしくはその情報を説明するための文字数の多さが動作変更の有無に影響していると考えた．

今後の課題として，アーム型ロボット本体の情報とその情報を説明するための文字数の多さのどちらに原因があるか調べる必要があると考え，追加実験を行う．

参考文献

[1] 横小路泰義, 植村渉ら,”World Robot Summit 2018 ものづくりカテゴリー競技「製品組立チャレンジ」の概要, ” 日本ロボット学会誌, 2019, 37 巻, 3 号, pp.208 – 217.

[2] A. Vaswani, et al.,“ Attention is All you Need, ” Neural Information Processing Systems, pp. 6000 – 6010, 2017.

〒 162 東京都新宿区津久戸町 4-7 OS ビル 402 号室 03-5261-3401 Fax: 03-5261-3402

(本研究会についてのお問い合わせは下記にお願いします.)

SIG-AI-Challenges web page; `https://www.osaka-kyoiku.ac.jp/~challeng/`