F-023

深層学習を用いたテニス打球コースの実時間予測手法の検討
A Study on Real-Time Prediction of Tennis Shot Trajectories Using Deep Learning Using Pose Estimation Data

岩田 雄介† 田村 仁† 大久保 友幸†
Yusuke Iwata Hitoshi Tamura Tomoyuki Okubo

1. はじめに

ロボティクス技術の進化により，人間の動作を予測して適切に反応するロボットが求められている．特に，駆動や処理に時間がかかるため，事前に動作を決定することが重要である．

テニスロボットを例に考える．ボールの速度とロボットがその着弾地点に行く速度が同じと想定した場合，相手が打ったボールの弾道を予測する時間やカメラの遅延があるので，相手が打った瞬間にロボットが動けないので，返球することができない（図 1）．

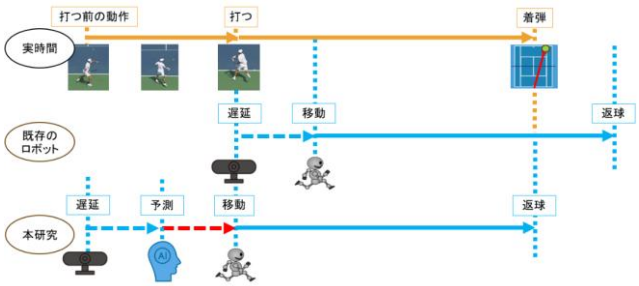

図 1　時間の関係

そのため，本研究では，深層学習を用いて動作予測を行い，テニスロボットが返球可能なアルゴリズムの開発と，シングルボードコンピュータ上でモデルを動作させることによる実運用に適したシステムの構築を目指す．

2. 先行研究

本研究では，想定するラリー動作が現実的に実現可能であるかを評価するために，ロボット・人間・ボールの移動速度に関する先行研究[1]を調査した．

ロボットの移動性能に関しては，車椅子型移動機構にアームロボットを搭載し，ラケットを持たせて自律的にラリーが可能なロボット「ESTHER」[2]が知られている．また，ヒューマノイド型でありながら二次開発が可能な「Unitree H1」[3]も注目すべき機体である．本研究ではこの 2 機種を対象とし，それぞれの最大移動速度および，コート中央から端まで移動するのに要する時間（以下，移動時間）について整理を行った．これらの情報は表 1 にまとめて示す．

表 1　動作速度と移動時間

	ESTHER	Unitree H1
速度(m/s)	4.34	3.3
加速(m/s^2)	1.42	不明
移動時間(s)	2.41	1.247

† Nippon Institute of Technology Graduate school Mechanical

一方，実際のラリーにおけるボールの飛行時間については，プロテニス選手である錦織圭選手とステファノス・チチパス選手の試合映像をフレーム単位で解析し，相手コートにボールが到達するまでの所要時間を測定した．その結果の一例として，飛行時間が約 1.28 秒であることが確認された．

これらのデータを表 1 に基づいて比較すると，「Unitree H1」の移動性能であれば，ボールが相手に打ち出された瞬間にその軌道を予測して移動を開始することで，現実的にラリーを成立させることが可能であると判断できる．

3. 関連研究

テニスにおける予測技術の研究は，ボールのトラッキング，打球フォームの解析，ショットコースの予測など，多岐にわたって行われている．しかし，対人ラリーを前提としたリアルタイム予測モデルに関しては，明確な事例は見られない．

例えば，テニス選手の歩行分析を用いてショットの種類をリアルタイムで予測するシステムが提案されているが，この手法はロボットを想定していないため，処理遅延などの実時間性の課題が考慮されていない[4]．「リアルタイム予測」と銘打ってはいるものの，処理全体における遅延時間や，実際に何秒先を予測しているのかといった定量的な情報は明示されていない．

さらに，テニスにおける次のショットの「位置」と「種類」を予測するために，記憶増強型深層生成モデル（Memory-augmented Deep Generative Model）を用いた研究も存在するが，こちらもリアルタイム性には課題がある[5]．

これらの先行研究を踏まえ，本研究では，対人ラリーを想定した実時間予測に重点を置いた手法の構築と検証を行う．

4. 提案手法

本研究では，ロボットがコート中央に構えた状態から，相手プレイヤーの動作と打球タイミングを捉え，ボールの進行方向（クロスまたはストレート）を数フレーム先まで予測し，それに応じてリアルタイムで移動動作を開始する手法を提案する．

まず，相手プレイヤーとボールの位置情報を高精度でトラッキングする必要があるため，人のトラッキングには YOLOv8[6]，ボールのトラッキングには TrackNet[7]を用いる．

YOLOv8 は，リアルタイム物体検出に特化した高精度なモデルであり，各フレームごとに人物の位置を高速かつ正確に検出可能であるため，プレイヤーの動作特徴抽出に適している．

TrackNet は，連続フレームを入力としてボールの軌道を時系列的に推定する CNN ベースのモデルであり，スポーツにおける高速移動物体の追跡に優れている．

これら 2 つの手法により得られた座標情報（CSV ファイル）を LSTM（Long Short-Term Memory）[8]に入力し，将来の数フレームにおけるボールの座標を予測する．LSTM は，時系列データにおける長期的な依存関係を学習可能であり，プレイヤーの動作やボールの位置変化などの時間的文脈を考慮することで，高精度な予測が可能となる．

本手法により，ロボットが人の打球意図を早期に察知し，実時間内に適切な位置へ移動できる環境の構築を目指す．

5. 予備実験

実時間での予測を実現するには，システム全体における各処理の遅延時間を正確に把握する必要がある．そこで予備実験として，カメラの遅延，物体検出およびトラッキング処理（YOLO・TrackNet）の処理時間，ならびに LSTM モデルによる予測処理時間の測定を行った．

5.1 カメラの遅延

カメラの撮影遅延については，先行研究の報告よりおおよそ 0.0021s であることが確認されている．

5.2 YOLO・TrackNet の処理時間

YOLO および TrackNet による物体検出および追跡処理に要する時間を測定した結果，1 フレームあたり約 0.0267s で CSV ファイルが出力されることが分かった．

5.3 LSTM モデルの推論時間

LSTM モデルによる 1 フレームの予測にかかる時間は，以下の設定で約 0.058s であった．この値を処理時間の参考値として採用する．学習条件は表 2 に示す．

Input，output は人の座標とボールの座標を計 6 データ（x，y を 3 データ）取った値である．

表 2　学習条件

n_seq	3
input_size	6
hidden_size	128
output_size	6
num_layers	3
batch_size	32
num_epochs	100
learning_rate	0.005

5.4 時間遅延の総合的評価

また，LSTM モデルへの入力フレーム数を変更した場合に掛かる処理時間をまとめた表を表 3 で示す．

表 3　LSTM の各フレームの処理時間

枚数（入力フレーム）	時間（s）
1	0.058
3	0.174
5	0.29
7	0.406

更に，カメラの遅延，YOLO および TrackNet による処理時間，ならびに LSTM モデルの推論時間を加えた総処理時間と，それを 25fps（1 フレーム＝0.04 秒）換算で表したフレーム数を，各入力フレーム数に計算し，表 4 に示す．

表 4　総処理時間とそのフレーム数

枚数（入力フレーム）	総処理時間（s）	フレーム換算
1	0.087	2.18
3	0.14	3.51
5	0.194	4.85
7	0.247	6.18

これにより，どれだけ先のフレームを予測する必要があるかを定量的に把握することができた．

5.5 データの前処理

表 4 より，各処理の遅延時間が明らかとなったため，それに応じて予測フレーム先のタイミングに合わせて画像データを調整した．

たとえば，入力が 3 フレームの場合，LSTM の処理遅延も考慮し，合計 7 フレーム分（入力 3 フレーム + 処理時間相当の 4 フレーム）を impact（打球）タイミングから遡って取得する必要がある．これにより，実時間処理に対応した予測モデルの構築が可能となる．を impact の瞬間からさかのぼって画像を保存する必要がある（図 2）．

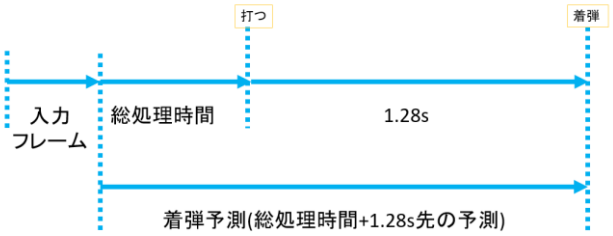

図 2　データ前処理

6. 実験準備

6.1 データ数

本研究では，TrackNet に含まれる既存のデータセットを使用する．このデータセットは 10 試合分のラリー映像を基に作成された画像群で構成されており，その中から 1 試合につき 1 ラリー（計 10 ラリー）を抽出して使用した．

深層学習におけるデータセットとしては十分な量とは言えないため，汎化性能を向上させる目的で K-Fold 交差検

証（Cross Validation）を採用し，限られたデータを有効活用する構成とした．

7. 実験

本実験では，入力データのフレーム数を表 4 に従って変更して実験を行う．

3，5，7 フレームを入力とするように変化させ，精度を確認していく．

7.1 座標化

YOLO・TrackNet を使用し，画像データから座標データに変換する．この時，外れ値や欠損地の補完をする必要がある．

外れ値の除去では，時系列的外れ値除去，中央値絶対偏差（MAD）による外れ値除去，標準偏差による外れ値除去を行っている．

欠損補完では，線形補完と前後補完を使用し，データを処理している．座標化した値をプロットした図を図 3 に示す．

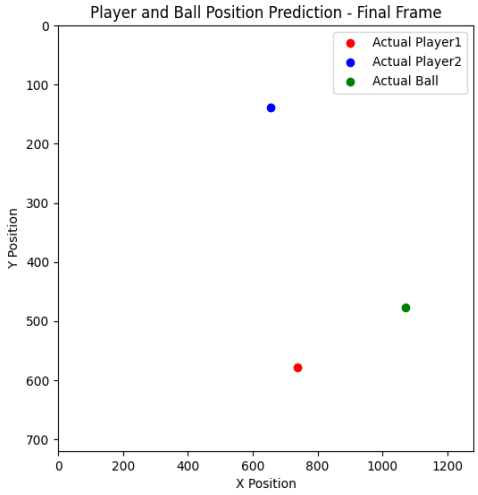

図 3　座標化した値をプロット

7.2 学習

LSTM（Long Short-Term Memory）を用いてラリー中の選手およびボールの動きを予測する深層学習モデルを構築した．入力データとして，フレーム単位で記録された 2 人の選手およびボールの位置情報（X, Y 座標）を使用し，一定数の時系列データ（n_seq=3, 5, 7）を入力として次のフレームの座標を出力するモデルを訓練した．学習条件を表 5 に示す．

学習データは複数のグループに分けられた CSV 形式の位置情報から構成され，前処理として線形補間および正規化（Min-Max スケーリング）を行った．K-Fold 交差検証（最大 5 分割）を用いてモデルの汎化性能を高める構成とし，各 Fold について学習を行った．モデル構造は，3 層の LSTM ユニットと 1 層の全結合層から構成され，損失関数に平均二乗誤差（MSE）を用いて最適化を行った．学習後，各 Fold のモデルは個別に保存された．

表 5　学習条件

n_seq	3, 5, 7
input_size	6
hidden_size	128
output_size	6
num_layers	3
batch_size	32
num_epochs	100
learning_rate	0.005

7.3 テスト

訓練済みモデルを用いて，未知のデータに対する逐次的な将来予測を行った．各グループに対し，直近 n_seq フレームの位置情報を初期値として与え，モデル出力を次の入力として繰り返すことで，フレーム単位で将来位置を 1 フレームずつ逐次的に将来の位置を生成していく．得られた予測値は正規化を逆変換し，Ground Truth との比較により精度を評価した．

評価指標として，各フレームにおける選手およびボールの平均二乗誤差（MSE）を算出し，特に最終フレームにおける予測精度を重点的に分析した．さらに，ボールの水平方向（左・右）を判定することで，予測の意味的正しさも確認した．可視化として，実際の画像上に予測結果と正解座標を重ねて描画し，出力画像と精度ログを保存した．最終的に，すべての Fold とグループにおける Ball の MSE を CSV 形式で集約した．

7.4 評価方法

評価方法はボールの座標を予測値と実測値の平均二乗誤差で評価する．数式は数式 1 に示す．

$$MSE = \frac{(\text{実測値} - \text{予測値})^2}{\text{データ数}} \quad (1)$$

8. 実験結果

実験結果をまとめた図を図 4 に示す．各入力フレーム数（n_seq = 3, 5, 7）における予測精度を比較したところ，入力フレーム数が 3 の場合に最も誤差が小さく，一方で 5 フレームを入力とした場合に最も誤差が大きくなる傾向が確認された．

この結果から，入力フレーム数が必ずしも多ければ高精度となるわけではなく，学習データの規模や対象となる動きの特性に応じて最適な入力長が存在する可能性が示唆された．

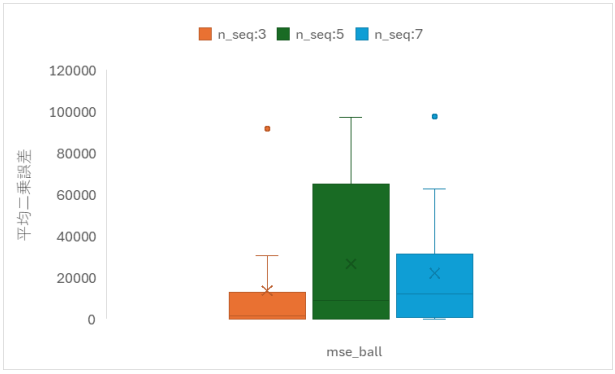

図 4　実験結果

値は平均二乗誤差の値なので，その値の平方根を取ってピクセルに変換したものを図 5 に示す．

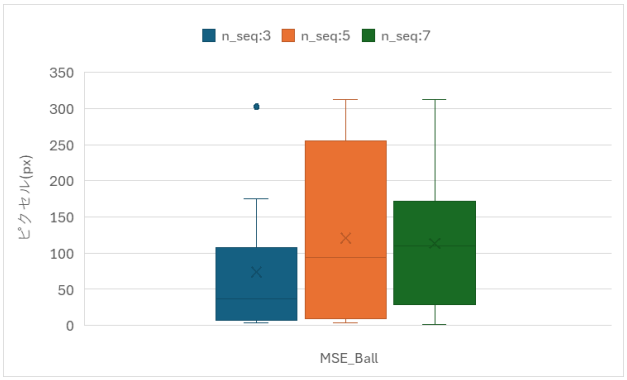

図 5　ピクセル変換

9. 考察

本実験において，入力 3 フレームが最も予測誤差が小さかった要因として，impact（打球時）との時間的距離が短いことが挙げられる．たとえば，入力 3 フレームと 7 フレームでは，6 フレーム＝約 0.28s の差があり，3 フレーム入力のほうが打球直前の情報をより濃く含んでいると考えられる．着弾に近いフレームを入力とするほど予測精度が向上するという結果は，時系列予測において「直近の情報」が重要であることを示唆している．

このことから，今後の展開としては 1〜2 フレームといった，さらに短い時系列を入力としたモデル設計により精度向上が期待できる．ただし，短い時系列のみではプレイヤーの動作特徴を十分に捉えきれない可能性があるため，骨格情報や姿勢の変化を捉える「姿勢推定モデル」の導入が有効と考えられる．

一方で，姿勢推定モデルは一般的に処理コストが高く，実時間予測の要件を満たすためには高速なモデルの選定や処理の最適化が不可欠である．今後の課題として，予測精度と処理速度のバランスを取るための技術的工夫が求められる．

10. まとめ

本研究では，テニスにおけるボールの進行方向を，打球直前の選手とボールの位置情報から予測する実時間モデルの構築を試みた．YOLOv8 および TrackNet を用いた座標抽出と，LSTM による時系列予測を組み合わせ，限られたデータセットでも高い予測性能を実現可能であることを示した．

特に，入力フレーム数が少ない方が誤差が小さいという結果は，リアルタイム応答が求められるロボットテニスシステムにおいて有益な知見である．また，将来的には姿勢推定の導入や高速な処理系の構築により，より高精度かつリアルタイム性に優れた予測システムの実現が期待される．

今後の展望としては，姿勢推定モデルの導入によってプレイヤーの動作特徴をより詳細に捉え，予測精度の向上を図るとともに，2 次元座標から 3 次元データへの拡張を行うことで，実際のロボットが返球動作を実行可能なシステムへと発展させていきたい．これにより，人と対戦可能な実用的なロボットテニスシステムの実現を目指す．

参考文献

[1] 岩田 雄介， 田村 仁， シングルボードコンピュータで動作できる手の動作予測モデルの作成と検討， 情報処理学会第 23 回情報技術フォーラム講演論文集 3 分冊，pp. 187-190(2024)

[2] Zulfiqar Zaidi， et al. ″Athletic Mobile Manipulator System for Robotic Wheelchair Tennis″. arXiv:2210.02517

[3] Unitree H1. https://www. unitree. com/h1. （最終閲覧日 2025/1/5)

[4] R. A et al.，″Deep Learning-Based Tennis Shot Type Prediction Using Gait Analysis,″ 2023 Intelligent Computing and Control for Engineering and Business Systems (ICCEBS)，Chennai，India，2023，pp. 1-4，doi: 10.1109/ICCEBS58601.2023.10449169.

[5] T. Fernando，S. Denman，S. Sridharan and C. Fookes，″Memory Augmented Deep Generative Models for Forecasting the Next Shot Location in Tennis,″ in IEEE Transactions on Knowledge and Data Engineering，vol. 32，no. 9，pp. 1785-1797，1 Sept. 2020，doi: 10.1109/TKDE.2019.2911507.

[6] Ultralytics，“YOLOv8 - Ultralytics,” 2023.https://github.com/ultralytics/ultralytics （参照日: 2025 年 3 月 115 日）

[7] Yu-Chuan Huang and I-No Liao，(2019) ″TrackNet: A Deep Learning Network for Tracking High-speed and Tiny Objects in Sports Applications,″ of arXiv，arXiv:1907.03698.

[8] Hochreiter， S. ， & Schmidhuber， J. (1997). ″Long Short-Term Memory. Neural Computation″， 9(8)，1735-1780. http://doi. org/10. 1162/neco. 1997. 9. 8. 1735