H-024

ニューラル場表現による大規模空間における自由視点映像生成手法の検討
A Method to Generate Free-Viewpoint Video for a Large-Scaled Scene Using Neural Radiance Fields

服部　航大† 謝　淳‡ 北原　格‡
Kota Hattori Chun Xie Itaru Kitahara

1. はじめに

ディジタル技術の進歩に伴うスポーツの情報化は，競技現場だけでなくスポーツ観戦形態にも変革をもたらしている．人工知能（AI）を用いた映像解析によって，試合の自動ハイライト生成や選手やボールの動作アニメーション作成など新たな視聴体験が提供されている．中でも，観戦者が自由に視点を切り替えながらの観戦を可能とする自由視点映像は，フィールド上や選手の足元など，実際にカメラを設置することが困難な視点から撮影されたような映像を見ることができるため，注目を集めている[1].

自由視点映像生成の分野では，複数視点で撮影した画像群にニューラル場表現を適用し，高精細なシーン表現を実現する研究が進んでいる．その代表例である NeRF [2]では，シーン全体を連続的な関数として扱い，3 次元座標と視線方向の入力によって与えられる光線において微分可能なボリュームレンダリングを適用し，その見え方を生成する．生成した画像と入力画像とのレンダリング誤差を最小化するようにシーンの放射輝度（色）と体積密度を学習する．

本研究で対象とするサッカーシーンでは，広大な空間をカバーするような撮影が行われるため，画像中で観測される選手領域は比較的小さくなる．このような画像群に基づいてニューラル場表現を学習する場合，スタジアムやフィールドといった背景領域が，学習過程における主要な最適化対象となり，選手領域の表現精度が低下するといった問題が発生する．また，広範囲を撮影するためには，撮影画角を広く（つまり焦点距離を短く）設定する必要があるが，そのような画像では，画面の端の方になると射影歪が生じるため，物体の見た目が大きく変化してしまい，ニューラル場表現の学習の妨げになることが懸念される．

大規模な屋外空間に存在する被写体の 3 次元形状を高速に復元する手法として，古山ら[3]は人物ビルボード法を提案した．人物ビルボード法は，選手位置に 2 次元平面（ビルボード）を配置し，多視点画像から切り出した選手領域（テクスチャ）を観察方向に応じて適宜マッピングする 3D モデリング手法である．具体的には，モデル化の対象となる被写体の 3 次元位置に人物と同等の大きさのビルボードを配置し，多視点画像から被写体領域のみを抽出した選手領域画像をビルボードに貼り付ける．ビルボードは常に観察者と正対するように回転させ，観察者の視点に応じて貼り付ける選手領域画像を切り替える．これにより，様々な視点から被写体を観察した自由視点映像を生成することが可能となる．一方で， 3 次元物体である選手の形状を平面近似するため，3 次元的な見え方の再現性や写実性に課題が残っている．

†筑波大学大学院 知能機能システム学位プログラム
Master’s Program in Intelligent and Mechanical Interaction Systems, University of Tsukuba
‡筑波大学 計算科学研究センター
Center for Computational Sciences, University of Tsukuba

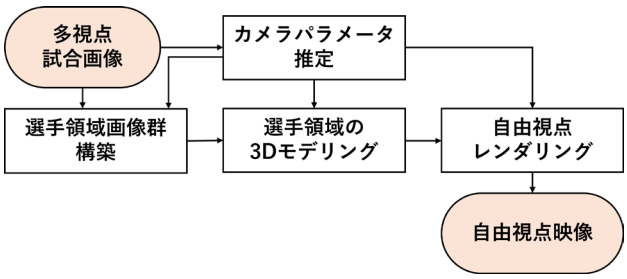

図 1　大規模空間における自由視点映像生成手法

本研究では，ニューラル場表現手法と人物ビルボード法の利点を統合し，大規模空間で実施されるスポーツイベントに適した自由視点映像の生成を目的とする．

2. 大規模空間における自由視点映像生成手法

提案手法の処理の流れを図 1 に示す．複数の視点でサッカーシーンを撮影した画像群を入力情報とする．前述した大規模空間におけるニューラル場表現手法の問題（小さい観測サイズと射影歪の影響）を解消するために，多視点画像から YOLO [4]などを用いて選手領域を切り出した選手領域画像群を生成する．各画像から切り出された選手領域は，カメラと選手間の距離が異なるため，画像上での観測サイズが変化し，ニューラル場表現の学習を難しくすることが考えられる．そこで，本研究では，人物ビルボード法を用いて観測サイズと射影歪みを補正した選手領域画像群を得る．それらの画像群に基づいてニューラル場表現を学習することで，大規模空間に適した自由視点映像生成を実現する．

2.1 多視点カメラによる撮影とカメラパラメータ推定

本研究では，ペナルティエリアをカバーする画角に設定した 31 台の多視点カメラでサッカーシーンを撮影する．その際，全てのカメラの光軸がペナルティキック地点で交わるように姿勢を設定する．

自由視点映像を生成するためには，多視点カメラの位置姿勢・焦点距離などのカメラパラメータが必要である．本研究では，Structure from Motion (SfM) [5]を利用し，多視点画像間の対応点情報から，カメラの位置姿勢，焦点距離を推定する．その結果得られる座標系（以下，SfM 座標系）において，フィールドのコーナーやセンターラインとタッチラインの交点などフィールド座標系での 3 次元座標が既知な点（ランドマーク）を三角測量する．さらに，上述した画像切り出しの影響を反映させた値を求め，それに基づきニューラル場表現を学習する．

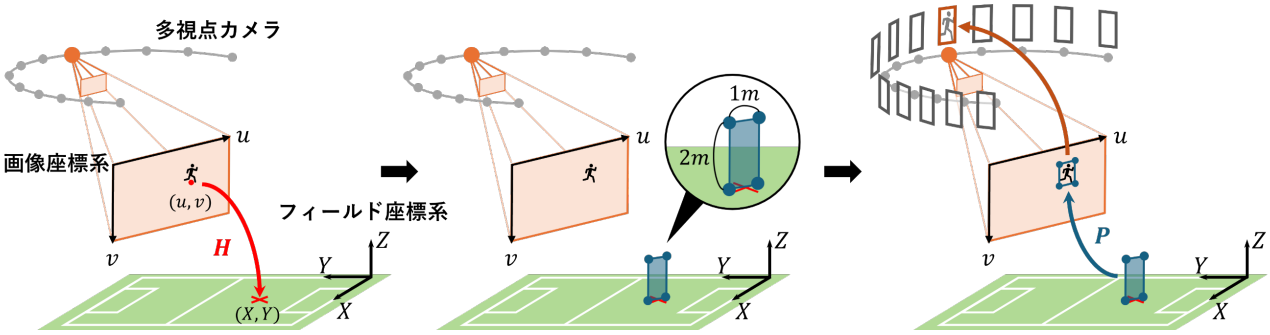

図 2　人物ビルボード法による選手領域画像の構築手法

2.2 座標系の変換

選手領域画像群の構築にあたり，実世界空間における座標系を設ける．図 2 に示すように，サッカーフィールドのあるコーナーを原点，ゴールラインを X 軸，コーナーマークと接するタッチラインを Y 軸，X 軸 Y 軸と直交する方向を Z 軸とする「フィールド座標系」を設定する． 2.1 節において SfM 座標系での三角測量で求めたランドマークの 3 次元座標と，それに対応するフィールド座標系における既知の 3 次元座標との対応関係に基づいて，座標系間の射影変換を計算する．

2.3 人物ビルボード法による選手領域画像群の構築

図 2 に人物ビルボード法を用いた選手領域画像群の構築の概要を示す．YOLO などの物体検出アルゴリズムを用いてサッカーシーンを撮影した画像から選手領域を検出する．本研究では，検出した選手領域矩形の下限中央を画像上での選手座標とする．画像座標系における中で観測されるランドマークとそれらに対応するフィールド座標系における位置座標より，画像座標系からフィールド座標系へのホモグラフィ変換行列Hを算出する．以降ではHを用いて，画像中の選手座標をフィールド座標系に変換する．

変換後の選手座標が下限中央となるように，フィールドに対して垂直で，かつ各カメラと正対する向きにビルボードを配置する．フィールド座標系におけるカメラ位置と選手座標から，カメラから選手に向かう方向ベクトルを算出する．この方向ベクトルに対する法線ベクトルの向きを，ビルボードの向きとする．ビルボードの大きさは，人物と同等の大きさである幅 1m，高さ 2m とし，選手座標を通過する法線ベクトル上で，選手座標から左右にそれぞれ 0.5m 離れた点と，それらから Z 軸方向に 2m 離れた点の座標を取得する．これらの三次元座標を画像座標系に射影し，射影された 4 点で囲まれた領域を選手領域画像として取得する．ここで，フィールド座標系から画像座標系への変換には，透視投影行列 P を用いる． 以上の一連の処理を，31 台のカメラで撮影された画像に対して適用し，選手領域画像群を構築する．

3. 自由視点映像の生成

本研究では自由視点映像の生成手法として，少数視点の画像から新規視点画像の生成が可能である pixelNeRF[6]を使用する．ニューラル場表現の代表例である NeRF は，数十枚から数百枚の多視点画像群が必要であるうえ，シーンごとに Multi-Layer Perceptron (MLP)を用いた最適化が必要である．一方で，pixelNeRF は大規模なデータセットによって事前学習された畳み込みニューラルネットワークを用いて，各入力画像から抽出した特徴マップを 3D サンプル点の位置情報と統合して MLP に入力する．これにより，シーンごとの再学習を行わずとも，少数視点から高精度な新規視点推定が可能になる．本研究では，この pixelNeRF の特性を生かし，31 台の限られたカメラ視点から得られる少数視点の選手領域画像を用いて，大規模なスポーツシーンにおける自由視点映像を生成する．

4. おわりに

本研究で扱うサッカーシーンのような大規模空間では，観測される選手領域のサイズが小さいうえ，視点ごとに観測サイズや射影歪が異なるという特徴がある．本稿では，そのような特徴を持つ多視点画像に対して，人物ビルボード法を用いて観測サイズと射影歪みを補正した状態で選手領域画像を取得し，それらに基づいたニューラル場表現の学習により，大規模空間に適した自由視点映像の生成手法を提案した．

参考文献

[1] キヤノン株式会社, “ボリュメトリックビデオシステム”, Canon Global, 2023, https://global.canon/ja/technology/volumetric-video2023.html
[2] Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R. and Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis, ECCV, pp. 405–421 (2020).
[3] T.Koyama, I.Kitahara, and Y.Ohta, “Live Mixed-Reality 3D Video in Soccer Stadium”, in Proc of IEEE/ACM Conference on ISMAR, pp.178-187 (2003)
[4] J.Redmon et al., ”You Only Look Once: Unified, Real-Time Object Detection”,CVPR (2016)
[5] Changchang Wu, “Towards Linear-time Incremental Structure from Motion”, International Conference on 3D Vision, pp.127-134, (2013)
[6] Yu, A., Ye, V., Tancik, M., & Kanazawa, A., “pixelNeRF: Neural radiance fields from one or few images.”, in Proc of the IEEE/CVF Conference on CVPR, pp. 4578–4587, (2021)