複合現実感研究会 MR2022-15
(2022.10. 6・ 7)

パフォーマンス撮影のための 演者立ち位置情報を基にしたカメラ位置姿勢生成手法

逸見 勲[*1]　　宍戸 英彦[*2]　　北原 格[*2]

Camera Position and Pose Generation Method Based on Performer's Standing Position for Performance Shooting

Isao Hemmi[*1], Hidehiko Shishido[*2] and Itaru Kitahara[*2]

Abstract --- The quality of the video is greatly affected by the camerawork. Unlike filming a movie, it is hard to reshoot live entertainment. Therefore, it is more important to consider the camera-work based on the information obtained in advance. The methods to generate camerawork using scenarios have been proposed, however, the methods have not considered changes in the standing positions of the performers on the stage yet. This study proposes an automatic camera work generation method that uses deep learning to output camera positions and postures for each scene based on the standing positions of the performers in each scene (keyframe) obtained from the stage script.

Keywords: Camerawork, Live Stage Performance, Deep Learning

1 はじめに

映像作品を制作するにあたり，カメラワークは作品を特徴づける要素の一つである．コロナウイルスの影響により，コンサートや舞台芸術などのライブパフォーマンスの分野において，ライブ配信やアーカイブ配信の市場は大きく拡大した．内閣府知的財産戦略本部によると，コロナウイルスによるイベント制限の緩和後も市場は拡大していくことが予想されている[1].ライブパフォーマンスの撮影は，映画やドラマと異なり撮り直しが困難であるため，公演前のカメラワークの検討がより一層重要になる．カメラワークを検討するにあたり，カメラワークは映像表現の一つであるため，最適なカメラワークを定義することは映像撮影技術の知識を持たない人には難易度が高く，映画やドラマの撮影においてもカメラワークは監督や演出家の感性によるところが多くなる．

本研究では，公演中のワンシーン（以下キーフレーム）における演者の立ち位置・姿勢情報を入力として，カメラの位置・姿勢を出力する深層学習を用いたカメラワーク生成手法の実現に取り組み，カメラワーク検討の補助を行うことを目指す．公演中の演者の立ち位置情報とカメラワークがセットとなったデータセットを作成する．また，深層学習を用いた手法として， Transformer [2]をベースとしたネットワークと Long Short Term Memory[3] (LSTM)をベースとしたネットワークの2種類を実装する．両ネットワークによって生成されたカメラワークの比較を行った結果を報告する．

2 関連研究

W.H.Bares ら[4]は，制約に基づくカメラワーク生成手法として，カメラショットの種類や配置に関する制約を設け，制約を満たすカメラの配置やアングルのパラメータを生成することで，カメラ位置姿勢の生成を行なった．しかし，カメラ位置姿勢を生成するには制約条件を詳細に記述する必要があることや，制約条件を解くだけでは映像撮影技術は考慮されず，撮影される映像は表現力に乏しくなるという課題が存在する．

井上ら[5]は，シナリオ情報に基づくカメラワーク生成手法として，オーケストラの演奏を対象に楽譜をシナリオ情報として用い，楽譜から得られる各演奏者の演奏機会から，撮影対象を選択するという手法を提案した．しかし，全ての演奏者が均等に撮影されることを前提として撮影対象を決定している．また，カメラ位置や演奏者の位置・姿勢が変化することは想定されていない．

本研究ではライブの進行台本をシナリオ情報として用い，公演中の演者立ち位置・姿勢とカメラの撮影位置・姿勢の関係を深層学習によって学習することで，カメラワーク生成に人手によるパラメータ設定を必要としない自動生成手法を検討する．

3 演者立ち位置に応じたカメラ位置姿勢生成手法

本研究では，公演中における演者のステージ上での位置姿勢情報とキーフレーム番号を時系列に入力することで，ライブパフォーマンスを撮影するためのカメラ位

*1 筑波大学システム情報工学研究群
*2 筑波大学計算科学研究センター
*1 Degree programs in Systems and Information Engineering, University of Tsukuba.
*2 Center for Computational Sciences, University of Tsukuba.

置姿勢情報（以下カメラワーク）を出力するネットワークを提案する．本稿では，Transformer ベースのネットワークと LSTM ベースのネットワークについて紹介する．

3.1 Transformer ベースネットワーク

Transformer をベースとしたネットワーク構造を図 1 に示す．Encoder へは，公演中の演者立ち位置・姿勢情報およびキーフレーム番号を，Decoder へは，対応するカメラワークを入力する．それぞれの入力は Fully Connected (FC)ブロックへ入力され，3層からなる FC 層によって段階的に 126 次元へと拡張される．その後， Positional Encoding によって時系列情報の埋め込み表現を付与された入力はそれぞれ Transformer Encoder と Transformer Decoder へ送られる．ここで，Transformer で用いられるマルチヘッドの数は3とし， Encoder，Decoder はそれぞれ6回同様の処理を繰り返す．Decoder からの出力は再び FC ブロックへ入力され，段階的に6次元へと圧縮されることでカメラワークを生成する．

3.2 LSTM ベースネットワーク

LSTM をベースとしたネットワーク構造を図 2 に示す．楽曲全体を通した演者の立ち位置特徴を得る LSTM ブロックと，カメラワーク生成を行う LSTM ブロックによって構成される．それぞれの LSTM ブロックは4層の LSTM 層によって構成されている．初めの LSTM ブロック（図左）に対し，公演中の演者位置姿勢情報とキーフレームを最終キーフレームまで時系列に入力する．最終キーフレームまで入力した LSTM 最終層から得られる Hidden state は楽曲全体の特徴を持つ．得られた Hidden state を各キーフレームにおける演者位置姿勢情報と結合させ，二つ目の LSTM ブロックへ時系列に入力することで，楽曲全体の特徴を踏まえたカメラワークの生成を行う．

3.3 損失関数

ネットワークによって生成されたカメラワークをターゲットとなる学習用教師データと比較し，損失関数を減少させるように学習を進める．本研究では，カメラの3次元位置に関する損失を $Loss_{pos}$，カメラの姿勢に関する $Loss_{angle}$ として，以下のように定義する．

$$Loss_{pos} = \Sigma_{i=0}^{n}\{(\hat{x}_i - x_i)^2 + (\hat{y}_i - y_i)^2 + (\hat{z}_i - z_i)^2\}$$

$$Loss_{angle} = \Sigma_{i=0}^{n}\{(\hat{q}_{0i} - q_{0i})^2 + (\hat{q}_{1i} - q_{1i})^2 + (\hat{q}_{2i} - q_{2i})^2 + (\hat{q}_{3i} - q_{3i})^2\}$$

ここで，x_i, y_i, z_i はそれぞれ，i 番目キーフレームにおけるカメラの3次元座標を表す．また，$q_{0i}, q_{1i}, q_{2i}, q_{3i}$ はそれぞれ，i 番目キーフレームにおけるカメラ姿勢オイラー角のクォータニオン表現を表す．ここで，クォータニオン Y 軸，X 軸，Z 軸の順番で回転するものとし，以下のよう

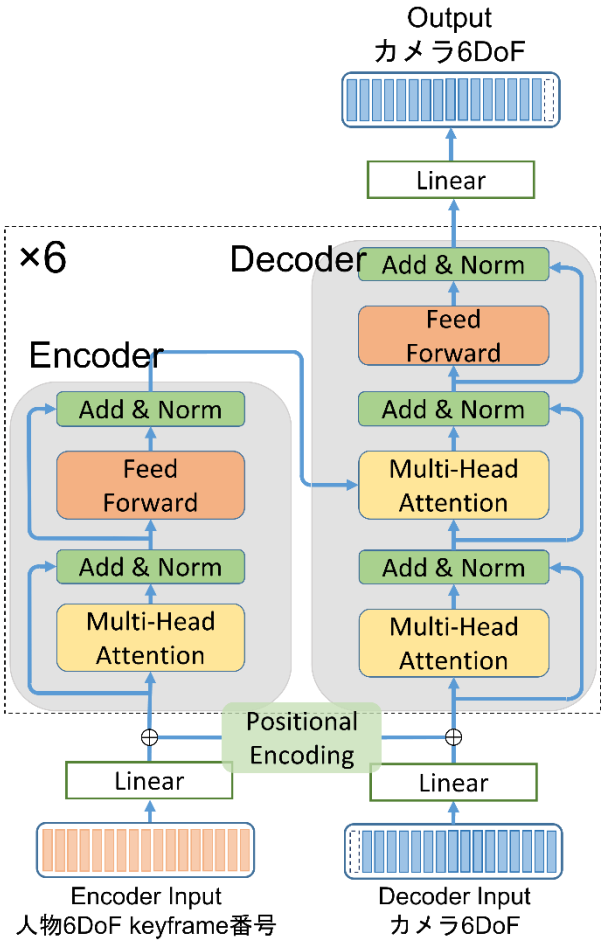

図 1 Transformer ベースネットワーク

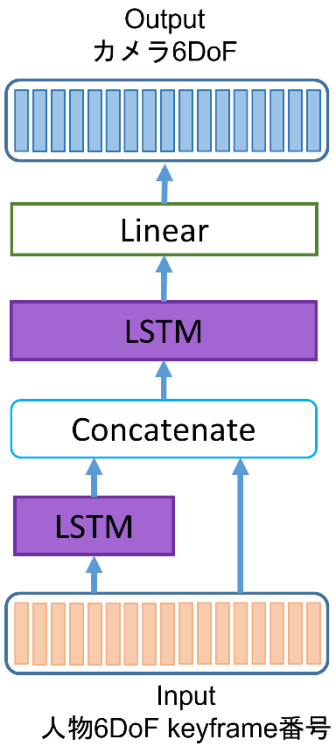

図 2 LSTM ベースネットワーク

に変換した.

$$\begin{pmatrix} q_0 \\ q_1 \\ q_2 \\ q_3 \end{pmatrix} = \begin{pmatrix} \cos\frac{\theta}{2}\sin\frac{\phi}{2}\sin\frac{\omega}{2} + \sin\frac{\theta}{2}\cos\frac{\phi}{2}\cos\frac{\omega}{2} \\ -\sin\frac{\theta}{2}\cos\frac{\phi}{2}\sin\frac{\omega}{2} + \cos\frac{\theta}{2}\sin\frac{\phi}{2}\cos\frac{\omega}{2} \\ \cos\frac{\theta}{2}\cos\frac{\phi}{2}\sin\frac{\omega}{2} - \sin\frac{\theta}{2}\sin\frac{\phi}{2}\cos\frac{\omega}{2} \\ \sin\frac{\theta}{2}\sin\frac{\phi}{2}\sin\frac{\omega}{2} + \cos\frac{\theta}{2}\cos\frac{\phi}{2}\cos\frac{\omega}{2} \end{pmatrix}$$

以上のように定義した損失を以下の式のように組み合わせることで$Loss_{total}$を求め，$Loss_{total}$が減少するように学習を進める.

$$Loss_{total} = Loss_{pos} + \alpha * Loss_{angle}$$

ここで，αは定数とする.

4 データセット

本研究では，データセットとして 3DCG ミュージックビデオ作製ソフトウェアである Miku Miku Dance (MMD)[6]のデータを利用する.

動画投稿サイトに投稿されている MMD 作品から，演者の立ち位置情報としてキャラクタモーションデータ，パフォーマンス撮影用のカメラワークとしてカメラモーションデータを利用した. キャラクタモーションよりキャラクタモデルの重心を表すセンタボーンの3次元座標を演者立ち位置情報として採用し，ステージ中央を原点としたステージ座標系座標のx, y, z，キャラクタモデルの3次元軸周りの回転θ, ϕ, ωを収集した. また，カメラモーションデータよりカメラ注視点のステージ座標系3次元座標x, y, z，カメラと注視点間の距離$dist$，注視点を中心とするカメラの回転θ, ϕ，カメラ座標系z軸周りの回転ωを収集した. カメラ座標系は注視点方向をz軸の正とする左手座標系で定義される.

4.1 データ前処理・拡張

収集したデータを利用するにあたり，データ間でのキャラクタとカメラの動きのスケールを統一するためにキャラクタモーションとカメラモーションのx, y, z座標，$dist$値の正規化を行う. また今回，利用したキャラクタモーションデータとカメラモーションデータ間で，キーフレームの位置および数が異なるため，キャラクタモーションデータとカメラモーションデータの各キーフレーム間について，配布 MMD モーションデータから得られるパラメータによって描かれるベジェ曲線を用いて補間することで楽曲中すべてのフレームのキャラクタモーションデータとカメラモーションデータを取得し，90 フレームごとの各モーションデータを取り出す.

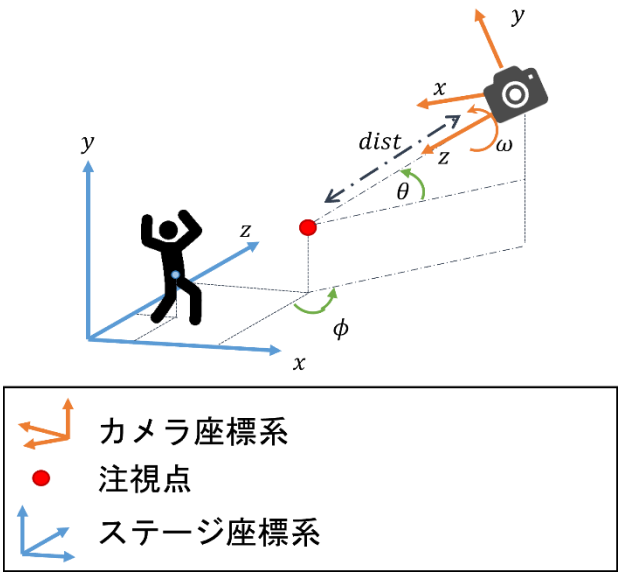

図 3 MMD 位置姿勢情報

本提案手法ではカメラの3次元位置及び姿勢を正解データとして利用するため，収集したカメラ注視点の3次元座標x^*, y^*, z^*，カメラと注視点間距離$dist$，注視点を中心とするカメラの回転θ, ϕを用いてカメラのステージ座標系3次元位置を以下の式のように求めた.

$$x = x^* + dist\cos\theta\sin\phi$$
$$y = y^* + dist\sin\theta$$
$$z = z^* - dist\cos\theta\cos\phi$$

また，得られた各モーションデータに対してデータ拡張を目的として，キャラクタモデルの位置姿勢，カメラの位置姿勢をステージ座標系$y-z$平面対称にしたデータを作成した.

5 評価実験

5.1 実験設定

LSTM ベースネットワーク，Transformer ベースネットワークともに，学習用教師データ 48 楽曲，評価用データ4楽曲を用い，実装したネットワークの評価を行った. 演者位置姿勢とキーフレーム番号からなる時系列データを与え，同時刻のカメラ位置姿勢が生成されるように学習を行う.

5.2 評価方法

本研究では,提案手法を用いて生成されたカメラワークを評価するために，テストデータから生成されたカメラワーク位置姿勢に関する損失を求める.また評価用データを用いて，キャラクタ位置姿勢および得られたカメラ位置姿勢座標を Unity 上にレンダリングすることでどのような映像が撮影されるのか確認を行うことで定性的な評価を行う.

5.3 実験結果

バッチサイズ2として学習を行い，損失関数に変化が見られなくなったところで学習を打ち切った．LSTM は 10000epoch，Transformer は 30000epoch でそれぞれ打ち切りとなった．テストデータから生成されたカメラワークのカメラ3次元位置および，カメラ姿勢に関する平均二乗誤差(MSE)は表1，表2のようになった．ここで，損失は各キーフレーム間における損失の平均となっている．

生成されたカメラワークは，カメラ3次元位置について楽曲ごとに手法間で大きく差が生まれた．カメラ3次元位置で差が生まれた一方，カメラ姿勢については手法間で大きな差はみられなかったことから，演者の立ち位置および姿勢を考慮して撮影が行えていることがわかる．しかし，評価時にカメラ3次元位置の誤差が大きくなるデータが存在することから，演者とカメラの距離感については学習が不十分であったと考えられる．

手法間における誤差が最も小さい Data4 について，カメラ位置を3次元上にプロットにした結果は図 4，Unity 上にレンダリングした結果は図 5 のようになった．図 4 を見ると，正解としているカメラワークはフレーム間で大きく移動し離散であるのに対し， 深層学習によって生成されたカメラワークは，フレーム間の移動が小さく，カメラの配置はひとつなぎの様になっている．また，図 5 を見ると Transformer によって生成されたカメラワークは，キーフレーム序盤ではカメラが大きく動き，演者を見上げるなどのカメラワークが生成された．しかし，中盤以降ではカメラの動きが小さくなり，演者をカメラの姿勢で追いかけるカメラワークが多くなった．LSTM では， カメラは演者の頭部正面に配置されることが多く，キーフレーム間での動きは少なくなっていた．

表 1 カメラ3次元位置誤差(MSE)

テストデータから生成されたカメラ3次元位置のキーフレーム平均二乗誤差

	LSTM	Transformer
Data1	0.12360	1.77922
Data2	0.14009	2.11219
Data3	1.90576	0.11856
Data4	0.13300	0.12212

表 2 カメラ姿勢誤差(MSE)

テストデータから生成されたカメラクォータニオン姿勢のキーフレーム平均二乗誤差

	LSTM	Transformer
Data1	0.627	0.635
Data2	0.265	0.178
Data3	0.623	0.620
Data4	0.167	0.157

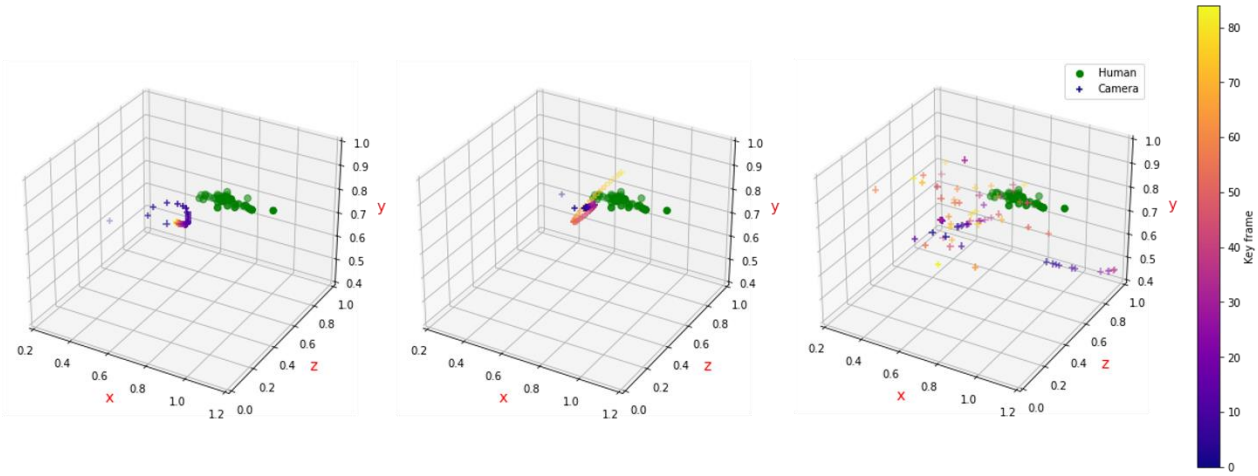

図 4 生成カメラ位置3次元プロット

左：Transformer 中央：LSTM 右：正解カメラワーク

図 5 同一キーフレームにおけるカメラワークの比較
左:Transformer 中央:LSTM 右:正解カメラワーク

6 おわりに

本研究では，ステージ上における演者の立ち位置情報から，撮影するためのカメラワークを生成する深層学習手法として，Transformer ベースと LSTM ベースの2種類を提案した．実験から，どちらのネットワークも演者の立ち位置姿勢情報を考慮してカメラの位置姿勢を生成していることが推察されたが，正解データのカメラ位置に比べ撮影位置の変化は大きくなく映像撮影技術を考慮したカメラワーク生成としては不十分であると考える．

今後はそれぞれのネットワークついて，フレーム間におけるカメラの移動量を損失に組み込むことで，より動きのあるカメラワークの生成を目指す．また今回，生成されたカメラワークの定量的な評価方法として MMD から収集したカメラワークのデータを正解カメラワークとして平均二乗誤差を求めたが，MMD から収集したカメラワークは必ずしも正解ではないと考えられる．今後も MMD から収集したカメラワークデータをネットワークの教師データとして利用するが， Unity 上にレンダリングされた生成カメラワークを用いて，SD 法によるアンケート調査を行うことを考える．定性評価方法を追加することで生成されたカメラワークの客観的な評価を行う．

参考文献

[1] 知的財産戦略本部:「デジタル時代のコンテンツ」戦略の方向性と課題の整理，第3回構想委員会配布資料; (2022)
[2] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser and Illia Polosukhin: Attention Is All You Need; Conference on Neural Information Processing Systems (2017)
[3] Hochreiter Sepp, Schmidhuber Jürgen.: Long short-term memory; Neural Comput. 1735–1780. (1997.11)
[4] William H. Bares, Somying Thainimit, Scott McDermott : “A Model for Constraint-Based Camera Planning”, Smart Graphics AAAI 2000 Spring Symposium (2000)
[5] 井上亮文，平石絢子，柴貞行，市村哲，重野寛，岡田謙一，松下温: シナリオ情報によるオーケストラ演奏のカメラワーク生成手法; 情報処理学会論文誌, vol(46), pp38-50 (2005)
[6] MikuMikuDance, https://sites.google.com/view/vpvp/