1ZL-05

# VR ダンスの自動撮影のためのカメラモーション生成

服部　嵩†　　長尾　確‡
名古屋大学　大学院情報学研究科†‡

## 1. はじめに

COVID-19 による影響を受け、ヘッドマウントディスプレイ(HMD)を着けて楽しむ VR コンテンツが注目されている。特にバーチャルライブやダンスパフォーマンス(VR ダンス)が注目されているが、多くの人が HMD を所持していないため、VR ダンスを 2D の映像コンテンツとして視聴する傾向にある。この実情を踏まえ、VR ダンスの映像化には専門的な撮影・編集技術が不可欠であり、演出には個々の演出家の感性が強く影響している現状がある。そこで本研究ではカメラモーションを自動生成することによって視聴者好みの映像を作成する手法を提案する。カメラモーションをカット割りとカメラワークの二つの要素に分け、それぞれ楽曲情報、ダンスのモーション情報から深層学習を用いて生成する。

## 2. カメラモーション生成手法

カメラモーション生成は高 FPS (Frames Per Second)でのカメラ位置生成といえる。これまでにカメラ位置の生成手法として様々な提案がなされてきた。W.H.Bares ら[1]は制約によってカメラ位置やアングルを決定する生成手法を提案した。しかし、この手法は制約条件を詳細に定義する必要があり、かつカメラの表現力が不足するという課題が存在する。逸見ら[2]はその課題を解決するために演者の位置情報を基に深層学習を用いてカメラの位置を生成する手法を提案した。しかし、この手法は 90 フレーム毎の位置生成でありダンスを撮影するのに十分な FPS であるとはいえない。

本研究では、ダンスのモーション情報からより高 FPS でのカメラ位置生成を行う。また、カット割りによってカメラ位置の連続性が損なわれることを考慮して、カメラモーション生成をカット割りの推定とカット間のカメラワーク生成の 2 つのタスクに分けて考える。全体的なカメラモーション生成の流れは図 1 のようになる。

Camera Motion Generation
for Automatic VR Dance Shooting
†HATTORI, Takashi (ch.53m.0682@s.thers.ac.jp)
‡NAGAO, Katashi (nagao@i.nagoya-u.ac.jp)
†‡Graduate School of Informatics, Nagoya University

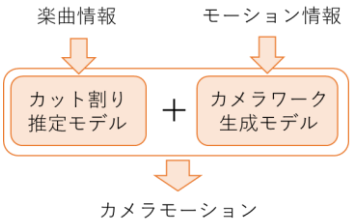


**図 1：カメラモーション生成の流れ**

## 3. データセット

本研究では、ミュージックビデオ作成用の 3DCG ソフトウェアである Miku Miku Dance (MMD)[3]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 作品からキャラクターモーションデータと撮影用のカメラモーションデータを取得して利用した。

### 3. 1 データの前処理

MMD のアバターモーションデータは 3D キャラモデルの規格である標準ボーンと準標準ボーンが入り混じっているため扱いが困難である。そのため、モーションをゲームエンジン Unity 上で再生し、アバターの標準ボーンの位置と角度を記録することによって統一的なデータへ変換する。モーション再生時には同一アバターを利用することによってアバターサイズの正規化も同時に行う。Unity で記録したデータはフレーム間の時間が不安定であるためスプライン補間を用いて 100fps のデータにリサンプリングした。

### 3. 2 カット割りラベルの作成

カット割りが起こるタイミングを判断するためにカメラの位置座標を利用した。フレーム間でのカメラの位置座標の移動距離が閾値を超えたタイミングをカット割りタイミングとして各フレームにラベルを付与した。

## 4. カット割り生成モデル

本研究では楽曲情報からカット割り生成を行う。楽曲中の内、特定のタイミングを推定する手法には C.Donahue ら[4]が提案した音響特徴量から教師あり学習を用いて特定のタイミングを推定する手法がある。しかし、奥村ら[5]は同様の手法を利用しても優れた精度が出せず、その理由を推定するタイミングの頻度の低さだとして

いる。我々の研究ではカット割りというタスクの性質上、C.Donahue らの場合と比較してイベントの発生頻度が低いことから同様の教師あり学習による手法は適していないと考えられる。

3.2 項にて作成したカット割りラベルを分析した結果、カット割りについては 1,2,4,8,16 カウントごとに行われることが多く、またサビの区間はサビ以外の区間に比べカット割りの平均数が約 1.1 倍多いことが分かった。このような特徴を基にトップダウンにカット割りを推定する手法として強化学習を利用する。カット割りの間隔と平均数を報酬として設計しカット割りタイミングの生成を目指す。

## 5. カメラワーク生成モデル

カメラワークの長さは固定長でない。一般的には最長の時間長にパディング処理することが多いが、長いモーションの生成タスクは難易度が高い。そこで、本研究ではカメラワークの動きは時間的に拡縮しても破綻が少ないと仮定する。カメラワークの長さを 4 秒に拡縮し、固定することによってタスクの難易度を低下させる。

学習モデルは Long Short Term Memory (LSTM) をベースとしたネットワーク構造であり、入力はダンスモーションである。入力を LSTM にて時系列的に処理し、Variational Autoencoder (VAE)デコーダによって潜在空間からカメラワークを生成する。カメラワーク生成モデルアーキテクチャは図 2 のようになる。

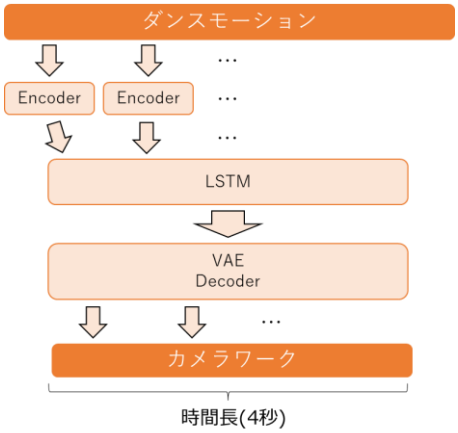


図 2：カメラワーク生成モデル

### 5. 1 学習用データの前処理

3.2 項にて作成したカット割りラベルを基にダンスモーション、カメラモーションを 1 カットごとに分割した。本研究では学習モデルの出力は時間長を固定し 4 秒とする。そのため、時間長が 4 秒となるようにカメラモーションに対して再度スプライン補完を用いたリサンプリングを行った。

### 5. 2 カメラワークの VAE

カメラワークの生成を行うために、モーションの生成タスクに良く用いられる VAE で学習した Decoder を使用する。カメラワークは各フレームについてカメラとカメラが向く方向の基準となる注視点それぞれの 3 次元座標を持つ 100fps のデータである。収集した 120 曲のデータの内、110 曲を訓練用、10 曲を評価用として使用した。

本研究では VAE の精度を向上させるため、収集したデータに追加してカメラモーションデータを作成し、学習モデルの事前学習を行った。収集したデータセットの位置座標の分布を基にカメラの移動開始位置と移動停止位置をランダムに決定し、その座標間を移動するモーションを多数作成した。MMD のモーションには座標間を移動する時の速度変化を調整するため、各フレームに対する座標の変化量を設定する補間曲線が存在する。そのため、座標間の移動時には多数の補間曲線を適応し、多様な動きを作成した。結果を比較した場合、表 1 に示すように事前学習ありの場合は、事前学習なしの場合に比べて高い精度を示した。そのため、事前学習ありのモデルをカメラワーク生成モデルに利用する。

表 1：カメラワークの VAE の精度比較

| | MAE | MSE |
|---|---|---|
| 事前学習なし | 0.244 | 0.253 |
| 事前学習あり | 0.124 | 0.078 |

## 6. まとめ

本稿では、楽曲情報とダンスのモーション情報からカメラモーションを生成する手法を提案した。次に、カット割りタイミングの分析からカット割りの傾向を示した。そして、カメラワーク生成の VAE を作成する時に事前学習を行うことにより、行わない場合に比べて高い精度を実現した。今後は、強化学習に基づくカット割りを推定するモデルを作成し、カメラワーク生成モデルと統合して、カメラモーション生成システムを完成させる予定である。

## 参考文献

[1] William H.Bares, Somying Thainimit and Scott D.McDermott, “A Model for Constraint-Based Camera Planning,” Smart Graphics AAAI 2000 Spring Symposium (2000)

[2] 逸見 勲, 宍戸 英彦, 北原 格, “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”, 複合現実感研究会 (2022)

[3] MikuMikuDance, https://sites.google.com/view/vpvp/

[4] Chris Donahue, Zachary C.Lipton and Julian McAuley, “Dance Dance Convolution,” Convolution. International Conference on Machine Learning (2017)

[5]奥村 綾,辻野 雄大, 山西 良典, “太鼓の達人の譜面難易度デザインに基づく楽曲中のキータイミングの推定に向けて”, エンタテインメントコンピューティングシンポジウム (2022)