複合現実感研究会 MR2022-16
(2022.10. 6・ 7)

ニューラル場表現と表情類似度に基づく 4次元ポートレート生成法

木村 文哉[*1]　　宍戸 英彦[*2]　　北原 格[*2]

4D portrait generation based on neural radiance field and facial expression similarity

Fumiya Kimura[*1], Hidehiko Shishido[*2] and Itaru Kitahara[*2]

Abstract --- This research aims to generate 4D portrait of a person that can be playback over a long time period. 4D portrait is a free-viewpoint video of a person with temporal changes in facial expression. In our proposed method, the parameters that represent facial expressions and head poses of a person are obtained from a video captured by a monocular RGB camera with continuously changing the viewpoint. Then, a neural radiance field (NeRF) is trained from the captured video and estimated parameters. Using this radiance field, 4D portrait is generated based on the similarity of the person's facial expressions. This paper describes a method for determining transition frames based on facial expression similarity to enable long playback.

Keywords: Free-viewpoint video, Neural radiance field, Expression, Head-pose, Video Textures

1 はじめに

本研究では，人物の4次元ポートレート生成法について述べる．4次元ポートレートとは，見え方が繰り返し再生される被写体を自由な視点から観察可能な映像である．カメラを連続的に移動させながら被写体（人物）を撮影した単眼 RGB 映像と，その映像から推定した表情パラメータを入力として，被写体の表情パラメータを含めた3次元モデルをニューラル場表現によって学習する．学習したニューラル場から，長時間に渡って表情が自然に変化し続ける人物の自由視点観察が可能な4次元ポートレートを生成する．

多視点映像を計算機内部で統合し，任意視点からの見え方を再現する自由視点映像の研究が活発に行われている[1][2]．従来の多視点映像の撮影法には，大きく二通りの方式が存在する．一つは，静止状態の被写体の周りを少数（通常は1）台のカメラを移動させながら多方向からの見え方を撮影する方法①，もう一方は，被写体を取り囲むように多数のカメラを配置し同期撮影する方法②である．前者の方法①で人物を撮影した場合，本来は動物体である人物が静止物体として自由視点映像中で観察されるため，生命感を感じ難いという問題が生じる．マネキンチャレンジで知られる Freeze Time Effect は，この効果を逆手に取って非現実感を演出する映像技法である．一方で我々は，生命感は映像の魅力を左右する重要な要素であると考え，その表現要因の一つである“動き”の提示機能を自由視点映像に付与することで，生命感の表現力を向上させる研究に取り組んでいる．後者の方法②で多視点映像を同期撮影すれば，動きのある自由視点映像の生成が可能であるが，高品質の自由視点映像を生成するためには大規模な撮影装置が必要となるため，日常シーンのポートレート撮影との親和性が低い．またカメラ台数および映像解像度の増加に伴い増大する映像データ量が，スマートフォンなどモバイル端末上での閲覧の障害となることも懸念される．本稿では，自然な状態の被写体の周りを移動する少数（通常は1）台のカメラで撮影した映像から，映像に含まれる動きを伴う自由視点映像を生成することによって上述した二つの問題の解消し，人物の4次元ポートレートを生成する手法を提案する．

被写体となる人物の動きには，骨格変動による大きな動きと，表情変化など体表面の微細な動きがある．本研究では，骨格変動による大きな動きではなく，テクスチャ変動により表現される体表面の微細な動きに着目する．例えば，じっとしている生き物でも，一定時間観察すると，呼吸や瞬きなどの動きによって生きていることを確認することができる．我々はこのような動きを提示することで映像に生命感を与えられると考えた．例えば，シネマグラフは，瞬きなど被写体の体表面の微小な動きによって生命感を表現し，観察者の注意を引き付ける映像技法であり，Web 広告などで人気が高まっている．本研究では，シネマグラフ表現のような，体表面の微細な動きを自由視点映像中の被写体に付与する．動きを有する被写体の周りを移動する1台のカメラで撮影した映像から，映像に含まれる動きを伴う自由視点映像を生成するためには，以下二つの研究課題を解消する必要がある．

*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.

研究課題1：一定長の映像から無限延長の映像を生成．
研究課題2：撮影時刻が異なる（非同期）多視点画像群からの被写体の3次元モデリング．

我々はこれまで，自由視点映像中の被写体に生命感を与える事を目的として，被写体の3次元形状に，微細な動きが含まれる映像をテクスチャとしてマッピングする手法を提案した[3]．マッピングするソース映像は一定長であるため，生成映像を長時間提示する際には繰り返し再生する必要がある．しかし同じ映像が繰り返されることによって映像観察者が見飽きてしまうことや，映像の最後から最初への切替り時に，被写体の見え方が大きく変わる問題が懸念された．そこで映像再生時にVideo Textures [4] を導入することにより，被写体の動きにランダム性を持たせて同一動作の繰り返しの発生を軽減すると共に，フレーム切替り時の見え方のギャップを低減することに成功した．本研究でも同様のアプローチにより研究課題1に対応する．

研究課題2には，深層学習を用いた3次元表現手法である Neural Radiance Field (NeRF)[5][6][7] の導入によって対応する．NeRF は，多視点画像群とカメラパラメータ情報を入力とした深層学習によって，高精細な3次元表現を実現する手法である．NeRF を用いた応用研究には，動的シーンの表現に対応した手法も存在する．特に，NeRFace [6] では，被写体の動的シーンを時刻パラメータではなく，ヘッドポーズパラメータと表情パラメータで表現する．異なる時刻でのシーンを付加情報である表情パラメータで記述することによって，NeRF の入力空間の範囲を絞ることができ，動的シーンの3次元モデリングが可能である．

しかし，上述した二つの対応法を単純に組みわせるだけでは，高品質な4次元ポートレート生成は達成困難である。なぜなら，研究課題1への対応は，撮影視点が変化しない被写体の映像を想定しており，一方で，研究課題2への対応は，多方向から撮影した撮影視点が変化する被写体の映像を想定しているからである．そこで我々は，図 1 に示すように，ヘッドポーズを正面向きに統一した映像から，画素値差分に基づく表情の類似度を算出し，遷移表情のペアを決定する．正面向きに統

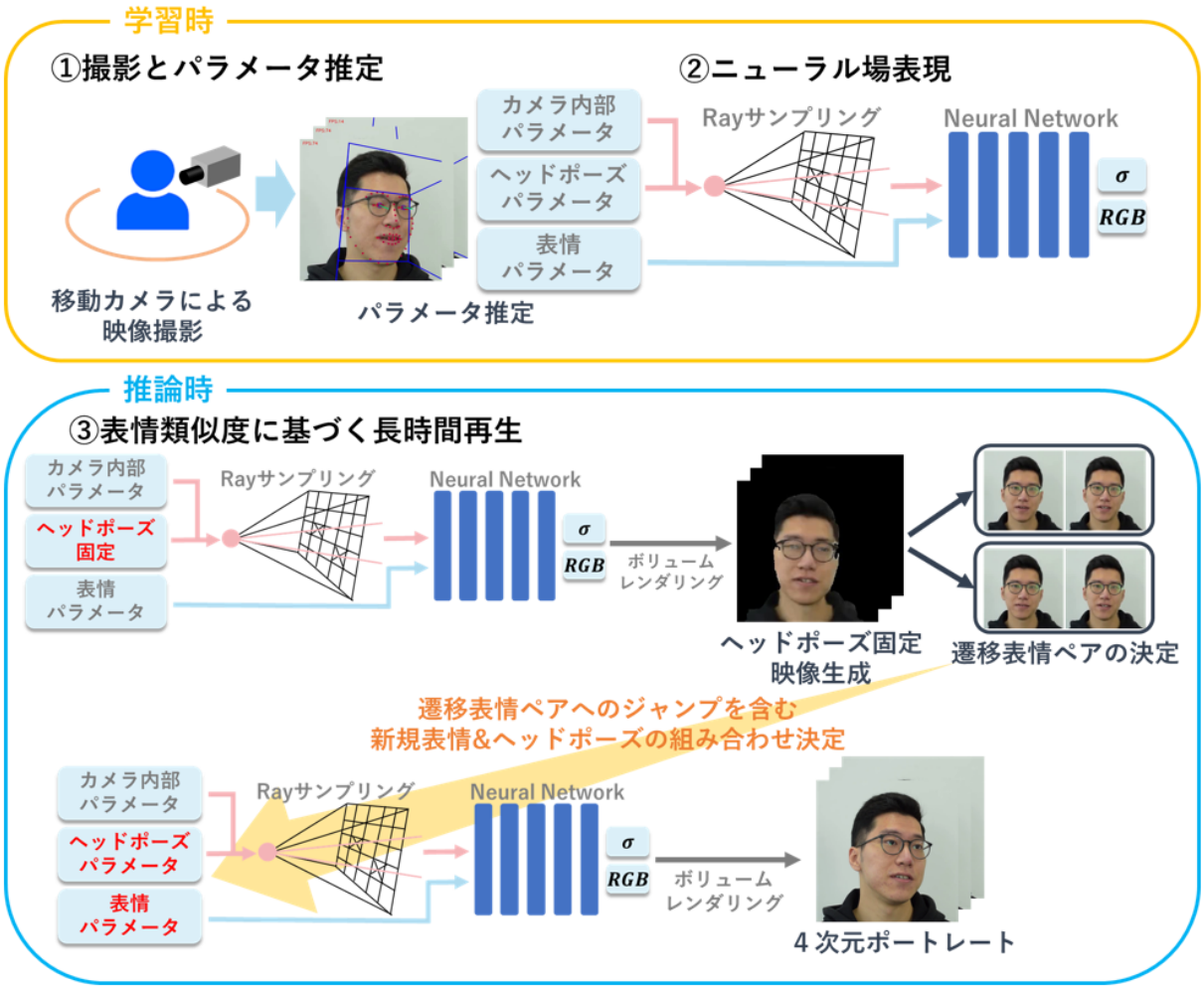

図 1　ニューラル場表現と表情類似度に基づく4次元ポートレート生成法
Fig.1 4D portrait generation based on neural radiance field and facial expression similarity

一した映像の生成では，撮影画像とレンダリング画像のヘッドポーズの変化量が大きい場合，レンダリング画像の画質劣化が懸念される．この課題の解決法については 6 節で後述する．

2 関連研究

2.1 動的な自由視点映像生成

自由視点映像は，多視点映像を計算機内部で統合することで撮影していない視点からの見え方を再現する技術である．Carranza ら[1]は，シルエット情報を用いて人物3次元モデルのモーション推定を行い，対応するテクスチャをマッピングすることで動的な自由視点映像を生成する手法を提案した．Collet ら[2]は，100 台以上のカメラを配置した多視点映像撮影システムを用いて，高品質でストリーミング可能な自由視点映像の生成手法を提案した．これらは，被写体の骨格の動きに焦点を当てた研究である．本研究では，自由視点映像中の被写体の骨格の動きを再現するのではなく，体表面の微細な動きを用いて生命感表現をする．

2.2 ビデオテクスチャの投影による微細な動きの表現

Structure-from-Motion (SfM)[8][9]によって多視点映像から生成した3次元モデルに対し，Video Textures[4]で生成したテクスチャを投影マッピング[10]することで被写体の体表面の微細な動きの再現が可能である．Video Textures は，短時間の映像内で類似フレームを検出し，ランダムフレームにジャンプさせることで，映像の連続性を維持しつつ単調な繰り返しではない長時間映像を合成する手法である．本研究でも Video Textures の考えに基づき，類似表情への遷移を発生させることで長時間映像の合成を行う．

我々が従来実現した手法では，微細な動きの表現に成功した一方で，多視点同期カメラを必要とするため，システム設置が容易でないという課題が存在した．本研究では，深層学習を用いることにより，単眼 RGB カメラ映像から動きを表現可能な自由視点映像の生成を行う．

2.3 深層学習を用いた3次元表現

Mildenhall ら[5]は，微分可能なボリュームレンダリングを介して3次元シーンを表現するネットワークを学習する Neural Radiance Field(NeRF)を提案した．NeRF では，同一時刻で撮影した多視点画像群を入力としてニューラル場表現の学習を行うことで，推論時に任意視点からの高精細なレンダリング画像生成を可能とした．現在NeRFを拡張した手法が多数提案されており，中には時間的な変化を実現する手法も存在している．Pumarola ら[7]は，NeRF を動的に変化するシーンに対応させ，多視点・多時刻での撮影画像の集合から，任意時間・任意視点画像の生成を行う D-NeRF を提案した．しかし，撮影対象となる物体は大きな動きを想定しているため，本研究で扱う表情変化のような微細な動きは表現が困難である．Gafni ら[6]は，事前に学習された顔追跡ネットワークにより予測された，被写体の表情やヘッドポーズを表すパラメータを NeRF の入力に組み込む NeRFace を提案した．結果として，単眼 RGB 映像から，時間的な動きを反映可能な3次元表現を実現した．本研究では，NeRFace に基づく微細な変化を含む3次元表現と，Video Textures に基づくフレーム遷移による長時間再生を統合することで，被写体の4次元ポートレートを生成する．

3 ニューラル場表現と表情類似度に基づく 4次元ポートレート生成法

本節では，単眼 RGB カメラで撮影した多視点画像から4次元アバターを復元し，表情類似度に基づき長時間映像再生を行う手法について述べる．

図 1 の①に示すように，被写体となる人物の周りを連続的に移動する単眼 RGB カメラ（内部パラメータは既知とする）で人物頭部を含む映像を撮影する．撮影した映像に顔追跡処理を施し，被写体の表情パラメータとヘッドポーズパラメータを推定する．図 1 の②では，撮影した映像，カメラの内部パラメータ，表情パラメータとヘッドポーズパラメータを入力としたニューラル場表現の深層学習を行う．学習したニューラル場表現に観察視点（レンダリング位置）の情報を与えることで，任意視点からの見え方が再現される．一定時間の映像から長時間の映像再生を実現するために，Video Textures と同様のアプローチを採用する．図 1 の③に示すように，学習したニューラル場表現とヘッドポーズパラメータを用いて正面から撮影した状態の見え方をレンダリングする（正面化）．正面化した画像群において表情類似度を算出し，その類似度に基づいてジャンプする表情（遷移表情ペア）を決定する．映像再生時に，遷移表情へのジャンプをランダムに繰り返すことにより，一定時間の入力映像から長時間の映像を再生する．本研究では，入力映像に含まれる被写体のヘッドポーズ動作と遷移ペアへの表情ジャンプを含む新規表情の組み合わせを決定し，適切に制御することで，4次元ポートレートを生成する．

4 映像撮影とパラメータ推定

連続的に移動する単眼 RGB カメラを用いて，被写体となる人物の頭部を一定時間撮影する．撮影においては，被写体の頭部が撮影画像に大きく映るようにカメラの光軸方向と位置を調整する．被写体の動きは，顔の向きの変化や表情の変化を含むものとする．

撮影映像の各フレームに顔追跡処理を適用し，表情パラメータとヘッドポーズパラメータを推定する．表情パラメータは顔追跡処理毎に異なる．Face2Face[11]では，主成分分析により求めた人間の顔空間を構成する 76 次元ベクトルで表情を表しており，OpenFace[12]では，Ekman ら[13]が定めた人間の表情を様々な顔面筋の動作の有無の組み合わせで表現する Action Unit を用いて表情を表している．本研究では，それぞれの顔面筋の微細な動きに着目し，OpenFace を用いて表情とヘッドポーズを推定する．表情とヘッドポーズは，それぞれ各フレームに対して，17 次元の Action Unit と 4×4 の行列で表現される．

5 深層学習による時間的変化を含む3次元表現

5.1 NeRFace

Neural Radiance Field(NeRF)とは，3次元空間のある点の位置(x,y,z)と視線ベクトル(θ,Φ)が入力された際に，その点の色(R,G,B)と密度 σ を出力するようなニューラルネットワークにより表現された場のことである．色と密度の情報からボリュームレンダリングにより，レンダリング画像を出力する．損失関数としては，レンダリング画像と実際の撮影画像の画素値の L1 誤差を使用する．NeRF では，視線ベクトルの存在により，視線の変化による反射特性の違いも表現することが可能である．応用研究である NeRFace では，入力映像中の各フレームの表情とヘッドポーズのパラメータを入力に付加することで，時系列的な変化を含む3次元表現を可能にしている．NeRFace では，カメラの移動の代わりに，ヘッドポーズの移動量を入力時に与えることで，画像を多視点化している．推論時には，表情とヘッドポーズのパラメータの組み合わせを変更することで，撮影時に存在しなかった画像のレンダリング（見え方の再現）が可能となる．

5.2 NeRFace の導入

本研究では NeRFcace のネットワーク構成において，入力の表情パラメータに OpenFace で推定した Action Units を用いる．学習時には，撮影に用いたカメラの内部パラメータ，および推定された各フレームの表情パラメータとヘッドポーズパラメータを入力データとして与え，深層学習を行う．損失関数には，正解画像とレンダリング画像の各ピクセルの平均二乗誤差を用いる．推論時には，表現したい表情パラメータとヘッドポーズパラメータを入力することで，自由視点画像を生成する．

Video Textures と同様のアプローチに従い被写体の表情が類似したフレームへのジャンプをランダムに実行しながら，NeRFace で任意の表情・姿勢の顔画像をレンダリングすることで，表情パラメータが自然に変化しつつ単調な繰り返しが生じにくい長時間自由視点映像が生成される．次節では，フレームジャンプの決定法について詳細に述べる．

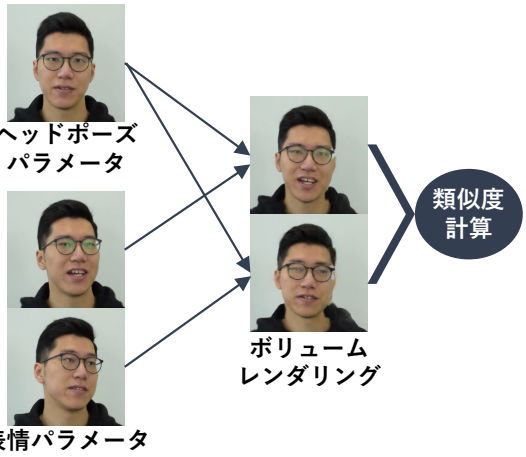

図 2 同一フレームのヘッドポーズパラメータを使用して被写体の見え方を統一，レンダリング画像間で類似度を計算

Fig2. Unify head poses of subjects using head pose parameters in the same frame and calculate the similarity between rendered images

6 表情類似度に基づく長時間再生法

6.1 表情類似度に基づく遷移フレームの決定

撮影した映像中の被写体の顔の向きは，時間変化に伴い様々に変化するため，従来の Video Textures で行っていた画素値の差分に基づくフレーム間類似度の算出は困難である．そこで本手法では，学習したニューラル場表現とヘッドポーズパラメータを用いて，図 2 右に示すように被写体を所定の視点から観察した見え方（例えば正面顔）をレンダリングし，その見え方に基づいてフレーム間類似度を算出する．

新規画像レンダリングの際，入力画像のヘッドポーズとレンダリング画像のヘッドポーズが大きく異なる場合，撮影画像中では観測されていない領域（耳など）の見え方を合成する必要があるため，レンダリング画像の品質が低下し，フレーム間類似度の計算に悪影響を与える．例えば，図 2 右では，撮影画像中で観測されていない被写体の左耳領域においてレンダリング精度が低下している．そこで，図 3 に示すように，レンダリング精度の高い領域のみを用いてフレーム間類似度を計算することで精度向上を図る．被写体画像群に背景差分などの領域分割処理を施し，共通して観測される領域（共通前景領域）を表すマスク画像を生成する．ヘッドポーズを統一した画像群にマスク処理を施した後， Video Textures と同様に画素値差分に基づくフレーム間類似度を計算し，遷移表情のペアを複数個選択する．

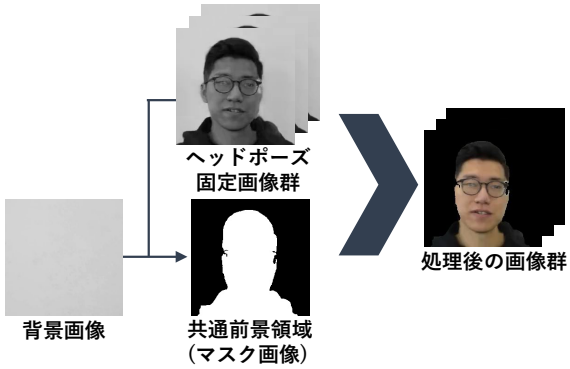

図 3 背景画像とヘッドポーズ固定画像群において
画素値差分により共通前景領域を抽出,
マスク処理で類似度算出に用いる領域を決定
Fig3. Extract common foreground area by pixel value difference between background image and fixed head pose image group, mask processing determines the area used for similarity calculation

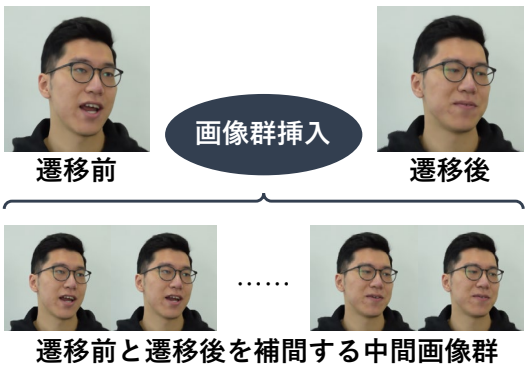

図 4 遷移前と遷移後の急な表情変化による
違和感を中間画像群の挿入で軽減
Fig4. Insertion of intermediate images reduces discomfort caused by sudden changes in facial expressions before and after transitions.

6.2 同一ヘッドポーズグループ内でのランダムな遷移

推定した遷移表情ペア情報に基づき, 一定確率で撮影時の時系列が後の表情から時系列が前の表情へのバックワードジャンプ(ランダムな表情ジャンプ)を発生させる. バックワードジャンプにより, 見え方がランダムに変化する長時間映像が生成される.

ランダムな表情にジャンプしながら新規画像をレンダリングする場合, 入力画像のヘッドポーズとレンダリング画像のヘッドポーズの差が大きくなると, 6.1 節と同様にレンダリング画像の品質低下が発生する. そこで, 撮影映像をヘッドポーズに基づいて分割(グループ化)し, レンダリング時のジャンプ先をヘッドポーズが類似した撮影画像グループを優先的に選択することにより, 画質低下を軽減する. ヘッドポーズのグループ化には, 撮影映像中の被写体の視線方向成分を用いる. 5節で推定したヘッドポーズを表す 4×4 行列から視線方向成分である3次元のベクトルをフレームごとに抽出する. 本手法では, 3次元ベクトルを入力として, k-means 法[14]により, 左向き, 正面向き, 右向きの三つのグループにヘッドポーズを分類する. 連続的に画像をレンダリングする中で, 異なるグループへとヘッドポーズが移行する場合, 表情パラメータも同一グループ内に遷移(強制的に表情をジャンプ)させることで, レンダリング画像の品質低下を防ぐ.

強制的な表情ジャンプでは, 遷移前後の1フレームで表情が大きく変化するため観察者が違和感を覚えることが考えられる. 図 4 に示すように, 強制的な表情ジャンプの発生時にモーフィングを導入し, 急な表情変化を防ぐことで観察者の違和感軽減を試みる. モーフィングとは, ある物体から別の物体へと連続的に変形させる画像表現のことである. Orita ら[15]は, 表情変化の時間が 0.2 秒程度であれば, 人は表情の変化を認知することができると述べている. 本研究では, 強制的な表情ジャンプの発生前後のフレームの表情パラメータから生成した中間画像群を遷移間へ挿入することにより, 0.2 秒以上の滑らかな表情遷移を生成する.

6.3 ヘッドポーズ動作の順再生・逆再生の組み合わせによる長時間映像生成

短時間の映像から長時間の無限再生可能な映像を生成するために, ヘッドポーズの動きにもランダム要素を組み込む. レンダリングの際に, ヘッドポーズの動きの順再生・逆再生をランダム確率で発生させる. ヘッドポーズの動きを決定した後, 6.1 節, 6.2 節で示すように表情の遷移順序を決定し, 単調な繰り返しでない長時間映像を合成する.

7 提案手法の実装

7.1 被写体の映像撮影と NeRFace の学習

連続的に移動する単眼 RGB カメラを用いて, 被写体となる人物の頭部を撮影し, NeRFace の学習を行った. 撮影機材はコンパクトデジタルスチルカメラ SONY RX0 を使用した. 撮影は, 1080 画素×1920 画素, フレームレート 60fps で 2 分間行い, その後 512 画素×512 画素へと映像をリサイズした. 7200 フレームの映像のうち 6000 フレームを Train データ, 200 フレームを Validation データ, 1000 フレームを Test データとして NeRFace の学習を行った.

7.2 表情類似度に基づく長時間映像生成

512 画素×512 画素, フレームレート 50fps で 120 秒間撮影した映像を用いて事前学習した NeRFace モデルを使用し, 表情類似度を推定した. 新規画像レンダリン

グには，学習映像と同様の条件で撮影した20秒間の映像のヘッドポーズパラメータと表情パラメータを用いた．レンダリング時の画像解像度は，256 画素×256 画素とした．

テスト映像のうち1フレーム目の被写体（正面向き）のヘッドポーズパラメータを用いて，所定の視点から観察した見え方をレンダリングした．続いて，ヘッドポーズ画像群と背景画像を用いて，共通前景領域（マスク画像）を抽出した．前景と背景の画素値差分を大きくするために，RGB画像のB成分のみを使用し，閾値を110として背景差分処理を適用した．マスク処理後の画像群から Video Textures の手法に基づき，19 組の類似表情ペアを推定した．それぞれの遷移表情ペアにおける，時系列が後の表情から時系列が前の表情へのバックワードジャンプの発生確率は 30%とした．テスト映像の入力映像が50fpsであるため，強制的な表情のジャンプ発生時には，表情パラメータの補間により生成した 9 枚の中間画像を挿入することで，0.2 秒間の表情遷移を発生させた．

ヘッドポーズ動作は，3%の確率で順再生と逆再生が切替わる設定とした．ヘッドポーズ動作と遷移ペアへの表情ジャンプを含む新規表情の組み合わせを決定し，連続的にレンダリングすることで，20 秒間のテスト映像から，40 秒間の新規映像を合成した．

8 おわりに

本研究では，単眼 RGB カメラによる被写体の撮影映像および，推定した表情とヘッドポーズパラメータからニューラル場表現の学習を行い，人物の表情類似度に基づく4次元ポートレートを生成する手法の提案を行った．特に人物の表情類似度に基づく長時間再生手法について述べた．

参考文献

[1] J. Carranza, C. Theobalt, M. A. Magnor, and H.-P. Seidel, “Free-Viewpoint Video of Human Actors,” ACM Transactions on Graphics, vol. 22, no. 3, pp. 569–577, 2003.

[2] A. Collet, M. Chuang, P. Sweeney, D. Gillett, D. Evseev, D. Calabrese, H. Hoppe, A. Kirk and S. Sullivan, “High-Quality Streamable Free-Viewpoint Video,” ACM Transactions on Graphics, vol. 34, no. 4, 2015.

[3] F. Kimura, H. Shishido, and I. Kitahara, “Lively Free-Viewpoint Video Generation Using Video Textures,” 7th IIEEJ International Conference on Image Electronics and Visual Computing (IEVC 2021), 2021.

[4] A. Schödl, R. Szeliski, D. H. Salesin, and I. Essa, “Video Textures,” Proceedings of SIGGRAPH, pp. 489–498, 2000.

[5] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, “NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis,” In Proceedings of the European Conference on Computer Vision (ECCV), 2020.

[6] G. Gafni, J. Thies, M. Zollhofer, and M. Niesner, “Dynamic Neural Radiance Fields for Monocular 4D Facial Avatar Reconstruction,” Proceedings of the IEEE/CVF Conference on Computer Visison and Pattern Recognition (CVPR), pp. 8645–8654, 2021.

[7] A. Pumarola, E. Corona, G. Pons-Moll, and F. Moreno-Noguer, “D-NeRF: Neural Radiance Fields for Dynamic Scenes,” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10318-10327, 2021.

[8] S. Agarwal, N. Snavely, I. Simon, S. Seitz, and R. Szeliski, “Building Rome in a Day,” 2009 IEEE 12th International Conference on Computer Vision (ICCV), pp. 72-79, 2009.

[9] J. L. Schönberger and J. Frahm, “Structure-from-Motion Revisited,” in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4104–4113, 2016.

[10] C. Everitt, “Projective texture mapping,” NVIDIA SDK White Paper, 2008.

[11] J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, and M. Nießner, “Face2Face: Real-time Face Capture and Reenactment of RGB Videos,” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2387-2395, 2016.

[12] T. Baltrušaitis, P. Robinson, and L.-P. Morency, “OpenFace: An open source facial behavior analysis toolkit,” in 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 1–10, 2016.

[13] P. Ekman and W. V Friesen, “Facial action coding system: a technique for the measurement of facial movement,” Consulting Psychologists Press, 1978.

[14] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu, “An efficient k-means clustering algorithm: analysis and implementation,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 7, pp. 881–892, 2002.

[15] A. ORITA, S. MUKAIDA, and T. KATO, “Perceiving a momentary change in facial expression,” The Japanese Journal of Cognitive Psychology, vol. 3, no. 1, pp. 1–11, 2005.