Vol.2020-CVIM-222 No.39
2020/5/15

深層学習による運動予測を用いた遠隔操作映像の時間補償

上田樹[1]　宍戸英彦[1]　北原格[1]

概要：本稿では，テレプレゼンス映像の伝送遅延の補償を目的として，映像中の人物の運動による見え方の変化を予測する手法を提案する．RGB-D カメラで取得した映像情報から被写体の骨格推定と身体形状（３次元点群）を生成する．深層学習を用いて推定した骨格情報から短時間後の挙動を予測し，予測した骨格形状に３次元点群を追従変形させることにより，運動に伴う見え方の変化を予測する．

キーワード：RGB-D 映像，深層学習，運動予測，３次元点群，自由視点映像

Time Compensation Method of Remote Operation Video Using Pose Prediction Based on Deep Learning

ITSUKI UEDA[1]　HIDEHIKO SHISHIDO[1]
ITARU KITAHARA[1]

1. はじめに

インターネット，VR 提示装置，遠隔ロボット操作技術の発展により，遠隔地で撮影した映像を用いて，その環境に仮想的に没入しながら操作を行うシステムの研究開発に注目が集まっている．そのような分野はテレプレゼンスやテレイグジスタンスと呼ばれている．遠隔操作の課題の一つが伝送遅延であり，過度な遅延時間は操作性の低下だけでなく，操作者の運動感覚と提示される視覚情報の間のずれによって身体動揺が誘発され，予期せぬ動作を引き起こす原因となることが指摘されている．テレイグジスタンスにおいて許容される誤差の検証 [1]によると，一般に遅延を認識できる閾値は 100~ 200 [ms]程度とされているが，身体動揺が誘発される遅延は 74 [ms]程度であるとしており，誤操作を防ぐにはより厳しい条件が求められる．特に，操作機器の周辺に人物が存在する環境では，人物との接触事故など伝送遅延による問題が深刻化しやすいため，伝送遅延補償処理による遠隔操作の安全性の向上が求められている．

人体の運動を予測し，その結果を提示することで伝送遅延の影響を軽減し，操作者の反応速度を改善した報告がある [2]. 我々は，その知見に着目し，人物の運動予測と自由視点映像技術 [3][4]を統合することで，一定時間経過したシーンの見え方を再現し，伝送遅延の問題を解消できると考えた．本稿では，図１に示すように，映像伝送に一定時間Δtを要する状況を想定し，その伝送遅延時間の補償を目的とした研究について述べる．操作対象物体は移動ロボットであり，操作者は，ロボットに設置したカメラで撮影した周辺映像を見ながらロボットを遠隔操作する. このとき，時刻tにおいて操作者が観察する映像は過去（時刻$(t-\Delta t)$）に撮影された映像である．Δtの時間補償処理によって時刻$(t-\Delta t+\Delta t=t)$の映像を生成し, 操作者に提示することで，操作者が映像を見て操作を行う時刻tと，ロボットが操作情報によって動作する時刻tを一致させる．

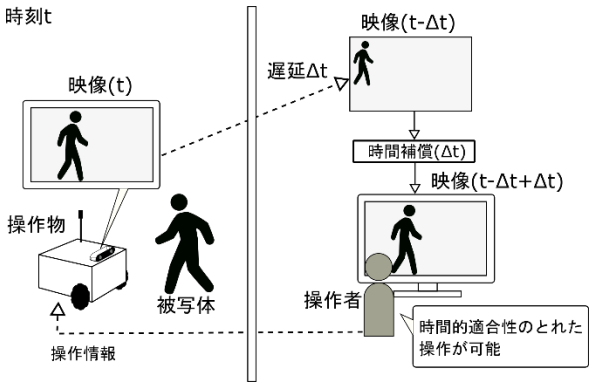

図 1 遠隔操作映像の時間補償

自由視点映像を生成するためには，注目物体の３次元形状情報とテクスチャ情報が必要である．本研究が撮影対象とする空間では，人物と遠隔操作する機器が混在するが，本稿では注目物体を人物とした場合の自由視点映像生成について考える．人体形状，またその挙動分析処理には，骨格情報が広く採用されている．画像情報から骨格を検出する代表的な手法の OpenPose [5]は単眼画像から２次元の骨格を高速かつ正確に検出することができる．しかし，３次元的な見え方の変化を再現する自由視点映像を生成するためには，３次元の形状情報が必要である．深層学習を利用することにより，２次元の骨格情報から３次元の骨格情報を推定する方法 [6]が提案されている．また，単眼 RGB 画像から奥行き情報を推定する手法 [7]により，２次元の骨格情報の３次元化も可能であろう．しかし，このように推

1 筑波大学
University of Tsukuba.

定された３次元情報では，世界座標系との幾何学的な関係の獲得が困難であるといった問題が残る．RGB-D カメラを用いて撮影することにより，可視光画像（RGB 画像）に加えカメラから被写体までの距離（奥行き画像）を得ることができる．近年では RGB-D カメラの低価格化が進み，広い用途での利用されている．本研究では，RGB-D 画像から３次元人物骨格情報を推定する．

本手法では，RGB-D 映像から推定した３次元骨格情報に深層学習を適用することで一定時間後の位置姿勢を予測し，その予測形状に基づいた見え方を自由視点映像技術で再現することにより，時間補償映像の生成を目指す．

2. 関連研究

2.1 人体の運動予測

３次元の運動予測は，キネマティクスを用いる方法と，深層学習を用いる方法がある．キネマティクスを用いる手法[8]では躍度最小モデルなどで人体の挙動のモデル化を行うが，誤差が蓄積しやすいため，高い精度で計測した入力情報が要求されることや，精度が保証できるのが比較的短い予測時間に限られるといった問題が存在する．深層学習を用いた運動予測法[2]では，人体運動に含まれる非線形性を学習するため，計測情報の変動にロバストであり，より長い予測時間を実現できる．

深層学習による時系列データのモデル化には，LSTM（Long short-term memory）などの RNN（Recurrent Neural Network）が用いられている [2]．複数のセンサから取得した情報を１チャンネルに束ねたデータに 1D-CNN（Convolutional Neural Network）を適用することで行動を認識する研究事例もある [9][10]．CNN は，モーションセンサのように力学的な意味を持つ時系列データとの親和性が高く，小規模なネットワークでも有効に働くため高速処理が可能であるという特長を有する．

先行研究 [2][10]では，画像内での各関節位置の２次元座標でモデル化し，入出力に用いられている．しかし，点群の移動予測へと活用するためには，各関節の位置を３次元で扱い，位置だけでなく姿勢の情報が必要になる．また直交座標系での表現では，関節間の距離や回転軸，可動範囲などの拘束条件の記述が困難である．本研究では，３次元空間での回転系の骨格モデルを構築し，関節角度を入出力とする畳み込みニューラルネットワークを構築する．

2.2 自由視点映像

移動するロボットの映像の時間補償を行うためには，異なる視点からの映像を生成する技術が要求される．このような問題に対し，映像から３次元形状を復元することで別の視点からの映像を再現する，自由視点映像技術が有用である．３次元形状復元としては，Structure-from-Motion [11]が知られている．静的な空間を対象とした場合は，カメラを移動させながら撮影した複数の画像間のマッチングを行うことで単眼映像からの形状復元が可能であるが，移動する人物のように撮影対象が時間とともに変形する場面では実行が困難である．また，深層学習により単眼 RGB 画像から奥行を推定する手法 [7]も提案されているが，世界座標系との幾何学的な関係の獲得が困難であるといった問題が存在する．ToF カメラやステレオカメラなどの機材を用いて撮影することにより，奥行情報をフレームごとに確実に獲得することができる．近年では RGB-D カメラの低価格化が進み， 自由視点映像への活用 [3][4]も多い．本研究では，RGB-D カメラの映像から自由視点映像を生成することにより運動予測映像を生成する．

3. 深層学習を用いた伝送遅延補償

提案手法のフローチャートを図２に示す．本システムは，「人物領域の移動予測」，「時間補償映像の生成」の２段階で構成される．

人物領域の移動予測においては，RGB-D カメラで人物を撮影し，奥行き情報から人物の３次元骨格情報の推定を行う．合わせて，人物領域のテクスチャ情報を RGB 画像から切り出す．推定した３次元骨格情報の時系列データを蓄積し，深層学習で生成した予測器を用いて３次元骨格の運動を予測する．また，RGB-D 映像から人物領域の３次元点群を生成する．予測した３次元元骨格の運動に合わせて３次元元点群を移動させることで，人物領域の移動予測を実現する．

時間補償映像の生成においては，遠隔操作ロボットのオドメトリを利用し，視点の移動予測を行う． ３次元点群から自由視点映像を生成し，予測した視点からの映像をレンダリングすることにより予測映像を生成し，伝送時間補償を実現する．

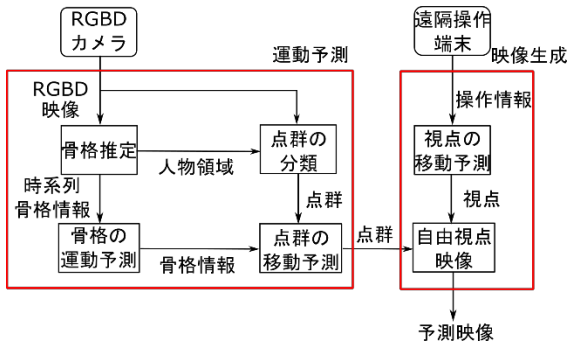

図２　提案手法の処理フロー

4. 点群の移動予測

4.1 骨格推定

4.1.1 回転系の骨格モデル

一般的なモーショントラッカーでは，各関節位置の３次

元直交座標を出力とすることが多い．しかし直交座標系ではリンクの長さの不変性や，各関節の回転軸と可動域などの拘束条件を記述することが困難である．そこで，骨格の運動を関節の角度で記述する，回転系の骨格モデルを構築する．

骨格モデルには，モーショントラッカーで取得可能な 16 点モデルを使用した．関節角度で表現するためには基準となる姿勢が必要である．Zero Neutral Position [12]を参考に，図３のような姿勢を基準ポーズに設定した．腰関節位置を原点，正面方向を x 軸，高さ方向を z 軸とする座標系をとり，各リンクのオフセット方向は XYZ 軸いずれかと平行となるよう設定する．

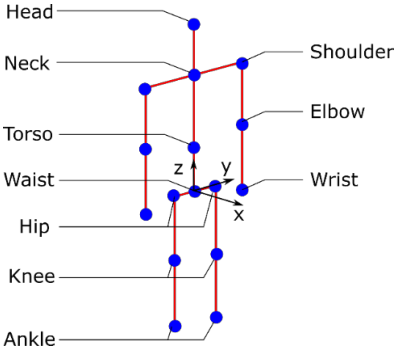

図 3　骨格モデルの基準ポーズ

各リンクでの変形は，Waist や Neck など一つの関節から複数のリンクへの分岐を表現できるようにするため，回転，オフセットの順に適用する．回転は直交座標系の XYZ 方向のいずれかを軸とし，リンクごとに設定された最大３回の回転を適用する．回転軸と可動域は，関節可動域表示ならびに測定法 [12]から，16 点モデルの頂点座標に影響しうる 15 項目 25 軸を表 1 に示すように設定する．

表 1 回転軸と可動域

関節名	左右	軸	可動域 1[°]	可動域 2[°]	可動域 3[°]
Waist	-	X	[-50, 50]		
Torso	-	YZ	[-30, 45]	[-40, 40]	
Neck	-	XY	[-50, 50]	[-60, 60]	
Neck	左	XZ	[-10, 20]	[-20, 20]	
	右		[-20, 10]	[-20, 20]	
Shoulder	左	YXZ	[-180, 50]	[-30, 135]	[-80, 60]
	右		[-50, 180]	[-135, 30]	[-60, 80]
Elbow	左	Y	[-135, 5]		
	右		[-135, 5]		
Hip	左	YXZ	[-125, 15]	[-20, 45]	[-45, 45]
	右		[-125, 15]	[-45, 20]	[-45, 45]
Knee	左	Y	[0, 130]		
	右		[0, 130]		

全身の位置と姿勢は，根幹となる Waist 関節で表現する．位置は３次元直交座標，姿勢は Z，X，Y 軸まわりの順での回転とする．

4.1.2 関節角度の計算

モーショントラッカーから得られた各関節の３次元直交座標から，回転系の骨格モデルでの各関節角度を計算する．

まず，根幹付近の関節の位置から全身の位置と姿勢を求める．Waist 関節はオフセットを持たないことから全身の位置と一致しなければならない．また Waist から Hip へのリンクは回転軸を持たないことから，右の Hip から左の Hip への方向ベクトルが Y 軸と一致するように全身の姿勢を決めるのが適切である．Waist と左右の Hip の直交系座標をそれぞれ$(p_x, p_y, p_z), (q_x, q_y, q_z), (r_x, r_y, r_z)$とおくと，全身の位置は$(p_x, p_y, p_z)$となり，姿勢は Z 軸回りの回転を$\phi$，X 軸回りの回転を$\psi$とおくと，式(1)(2)のように計算できる．

$$\phi = \mathrm{atan2}(r_x - q_x,\ q_y - r_y) \tag{1}$$

$$\psi = \mathrm{atan2}(q_z - p_z,\ \sqrt{(q_x - p_x)^2 + (q_y - p_y)^2}) \tag{2}$$

Waist から Torso へのリンクを YZ 平面上になるよう，全身の Y 軸まわりの回転を決める．Torso の直交系座標を(s_x, s_y, s_z)とおくと，Y 軸まわりの回転γは式(3)(4)(5)のように計算できる．

$$\gamma = \mathrm{atan2}(-s'_x,\ s'_z) \tag{3}$$

$$s'_x = \cos\phi(s_x - p_x) + cos\psi\sin\phi(s_y - p_y) + \sin\psi\sin\phi(s_z - p_z) \tag{4}$$

$$s'_z = -\sin\psi(s_y - p_y) + \cos\psi(s_z - p_z) \tag{5}$$

次に，関節の各回転軸の角度を求める．各関節の位置が分かっている場合，根元の関節から順番に角度を決定していくフォワードキネマティクスが知られている．しかし，回転軸の制限がかかっているとき，対象の関節だけでなく親関節まで辿っての調整を必要とするケースが存在する．例えば，図４の上段のように，Wrist の位置の誤差が回転軸と一致していない場合，下段のように親関節である Elbow の回転を調整する必要がある．そこで，Cyclic-Coordinate-Descent 法(CCD)による逆運動学の最適化を行う．

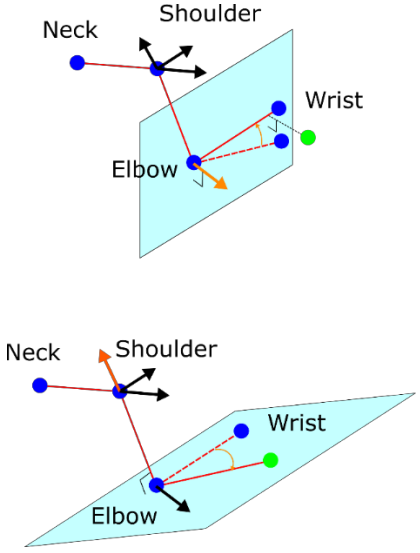

図 4　親関節に依存する回転

CCD では 1 回の計算ステップで一つの関節角度のみを最適化し，それをすべての関節に反復することで全体の近似解を求める．各計算ステップにおける最小化問題は，図 5 のようにリンクの根元から先端への方向が目標位置の方向と一致する関節角度を解に持つことから，１回の計算ステップ内では閉じた形式で一意に計算できる．

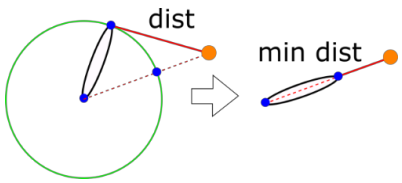

図 5　リンク回転量の最適化

回転軸に拘束のある 3 次元での回転の場合，図 6 のように計算できる．３次元基本ベクトルに対して，計算対象の回転より根本側の座標変換を適用することで，回転軸とそれに直交する平面が構成できる．平面上へリンク先端と目標位置を投影し，対象となる回転軸回りの回転量を求める．

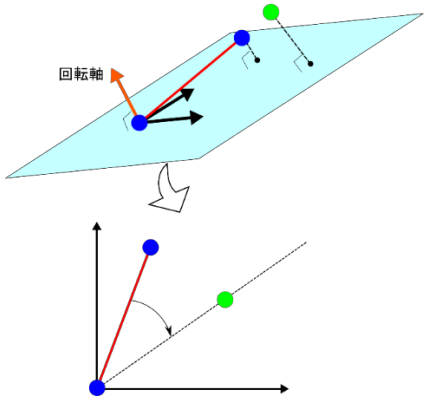

図 6　回転軸回りの回転量

これを根元から順に適用することで各関節の位置合わせが行われるが，回転軸が 2 軸以下の関節では目標位置とリンク方向が一致させられない場合がある．そこで，親関節までの複数のリンクで仮想的なリンクを構成し，親関節の関節角度の再計算を行う．回転量の再計算は回転順と逆順に行い，目標位置との誤差が閾値以下にならなければさらに親関節まで含めた仮想リンクを構成し，再計算を行う．これを全関節で繰り返すことで，関節角度が計算される．

4.2　骨格の運動予測

4.2.1 重心位置の正規化

本方式では，全身の位置と姿勢，各関節の角度を結合したものを入力に用いる．このうち全身の位置と姿勢は絶対座標系にて定義されており，値域が限定されていない．しかし周囲の状況が入力されない場合，地面と平行な平面内で全身の動作全体を回転または平行移動させても，区別されないことが望ましい．そこで，深層学習による予測器の前処理として，全身の位置と姿勢の正規化処理を行う．

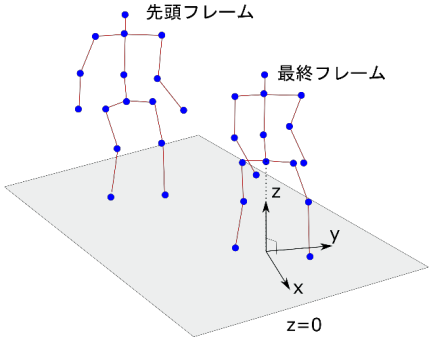

図 7　全身の位置と姿勢の正規化

図7に示すように，最終フレームの Waist 関節の位置を地上面へ投影した位置を原点，最終フレームの全身の Z 軸回りの回転が 0 となるような座標変換を行う．時系列データの最終フレームの Waist 関節の位置を(x_n, y_n, z_n)，Z 軸回りの回転角度をθとすると，座標変換M_l，絶対座標系への逆変換M_gは

$$M_l = \begin{bmatrix} cos\theta & 0 & -sin\theta & -x_n\,cos\theta + z_n sin\theta \\ 0 & 1 & 0 & 0 \\ sin\theta & 0 & cos\theta & -x_n sin\theta - z_n cos\theta \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad (6)$$

$$M_g = \begin{bmatrix} cos\theta & 0 & sin\theta & x_n \\ 0 & 1 & 0 & 0 \\ -sin\theta & 0 & cos\theta & z_n \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad (7)$$

と表される．深層学習への入力はM_lを用いて正規化し，予測器で推定された座標はM_gを用いて絶対座標系に変換する．

4.2.2 深層学習による予測器

計算した骨格モデルの状態について，３自由度の全身の位置をメートル単位で，２自由度の全身の姿勢と 25 自由度の各関節の角度をラジアン単位で表記して結合させた，31 次元のデータとして表現する．骨格モデルの状態の時系列データを入力とし，一定時間経過後の骨格モデルの状態を出力とする畳み込みニューラルネットワークを構成する．また入力する時系列データは 30fps で 64 フレーム使用する．

畳み込みニューラルネットワークにおいて，プーリング層を用いることで広い範囲を効率よく参照したネットワークを構築できることが知られている．一方で，時系列方向の分解能を下げるため，運動の高周波成分が失われて伝わってしまう問題がある．また，速度，加速度など力学的に有効な特徴量は浅い層で現れるが，ニューラルネットワークは非線形写像のため深い層まで伝達されにくい．このような問題に対し画像処理などの２次元畳み込み層を用いるニューラルネットワークでは，U-Net [13]や DenseNet [14] のように中間層の間で情報伝達を行う手法や，ResNet [15]

のようにショートカット構造を用いて残差の形にする手法などが考案されている．時系列データに対する１次元畳み込み層においても同様に，中間層からの情報伝達が有効であると考えた．時系列データの畳み込みでは，最新の特徴量は最終フレームに現れることから，入力，４，８層目の出力の最終フレームを畳み込みの最終層の出力に結合させる，図８のようなネットワークを構成する．

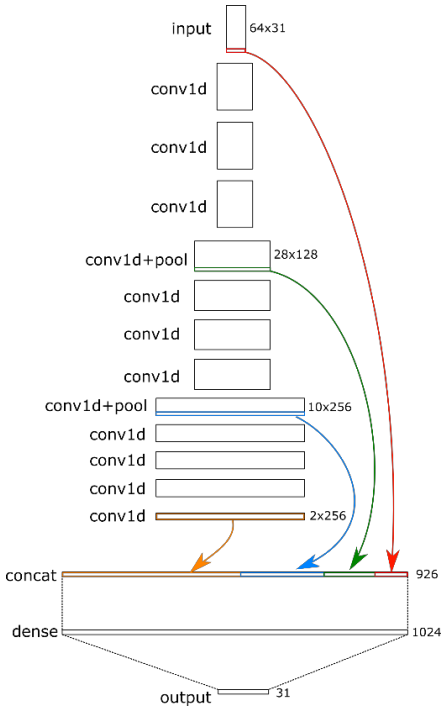

図８　運動予測のためのニューラルネットワーク

表２にネットワーク構成の詳細を示す．なお，Layer Type について，Co はカーネルサイズ 3 の 1D-Convolution，BN は Batch Normalization，MP は Max Pooling，FC は Full Connection を表す．

表２　図８のネットワークの構成

Layer Type	Output Channel	Frame Length
Input	30	64
Co+BN+ReLU	64	62
Co+BN+ReLU	64	60
Co+BN+ReLU	64	58
Co+BN+ReLU+MP	128	28
Co+BN+ReLU	128	26
Co+BN+ReLU	128	24
Co+BN+ReLU	128	22
Co+BN+ReLU+MP	256	10
Co+BN+ReLU	256	8
Co+BN+ReLU	256	6
Co+BN+ReLU	256	4
Co+BN+ReLU	256	2
Layer Type	Input Size	Output Size
Concat	-	926
FC+ReLU	926	1024
Output(Linear)	1024	30

ショートカット時に最終フレームのみを使用することで次元数を削減し，全結合層によるエンコードを安定させることができる．またショートカットした特徴量の結合には，U-Net に代表されるように次元数を合わせて順に加算する手法が一般的であるが，本手法では連結層を用いることで参照時間の違う特徴量を同時に扱えるような構造とした．

4.3　人物領域の３次元点群の移動予測

３次元骨格の運動予測に合わせて点群を移動させるためには，点群とリンクの結び付けが必要となるが，リンクから境界面までの距離は一定ではなく，例えば，腕を下した状態では胴体の一部が腕と紐づけられてしまう．そこで，図９に示すようにあらかじめリンクと結び付けた簡易的なメッシュモデルを作成し，メッシュモデルの面からの距離に応じて対応付けを行う．

人物領域の３次元点群について，各点群とメッシュとの距離を計算する．図 10 のように最近傍のメッシュから結び付けたリンクを求め，対応する座標変換を適用することで，運動予測に合わせた点群の移動を実現する．

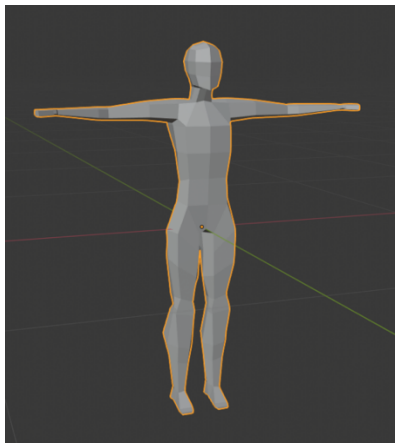

図９　人物の簡易メッシュモデル

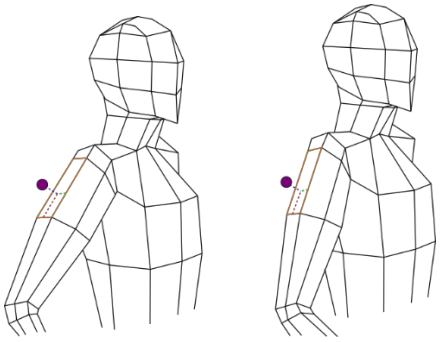

図 10　点群の対応付けと移動

5.　時間補償映像の生成

5.1　視点の移動予測

遠隔操作の操作情報は，操作者側のコンピュータに最新の

情報が記録される．操作対象のロボットの車輪やステアリングの物理的な情報を事前に取得しておけば，伝送遅延のある環境においてもオドメトリ [16]により最新の視点の位置を計算することができる．本稿ではロボットが独立２輪駆動であり，並進速度，角速度を操作項目としている場合を想定する．時刻 t の並進速度をv_t，角速度をω_tとすると，ロボットの位置$(x_t,\ y_t,\ \theta_t)$は式(8)(9)(10)のように計算できる．ここで，$(x_0,\ y_0,\ \theta_0)$はロボットの初期位置である．

$$x_t = \int_0^t v_\tau \cos(\theta_\tau) d\tau + x_0 \quad (8)$$

$$y_t = \int_0^t v_\tau \sin(\theta_\tau) d\tau + y_0 \quad (9)$$

$$\theta_t = \int_0^t \omega_\tau d\tau + \theta_0 \quad (10)$$

ただし，オドメトリによるロボットの位置推定は車輪の滑りを検出できず，誤差が時間とともに蓄積する．遅延の存在する環境であっても，SLAM などのセンサ情報を用いて誤差の蓄積を軽減することができる．時刻 s(s<t)における位置$(x_s,\ y_s,\ \theta_s)$が得られているとき，式(11)(12)(13)のように計算できる．

$$x_t = \int_s^t v_\tau \cos(\theta_\tau) d\tau + x_s \quad (11)$$

$$y_t = \int_s^t v_\tau \sin(\theta_\tau) d\tau + y_s \quad (12)$$

$$\theta_t = \int_s^t \omega_\tau d\tau + \theta_s \quad (13)$$

5.2 映像生成

5.2.1 メッシュ化

視点が移動した際，点群の密度が十分でないと生成した映像には欠損が生じてしまう．点群のメッシュ化を行うことで，分解能の不足した領域を補完することができる．点群からのメッシュ化ではポアソン方程式を用いた手法が知られているが，計算コストが大きく，撮影フレームごとにメッシュを生成するのは現実的でない．本研究では，映像の格子を利用してメッシュを生成する．

デプスカメラの映像では，各画素について奥行の情報が得られる．注目画素についてその右，右下，下に隣接する画素との奥行の差の絶対値を計算し，閾値以下のものを接続する．２本以上の接続関係があるとき，対応する３次元点群の頂点について，図 11 に示すように三角形メッシュを追加する．

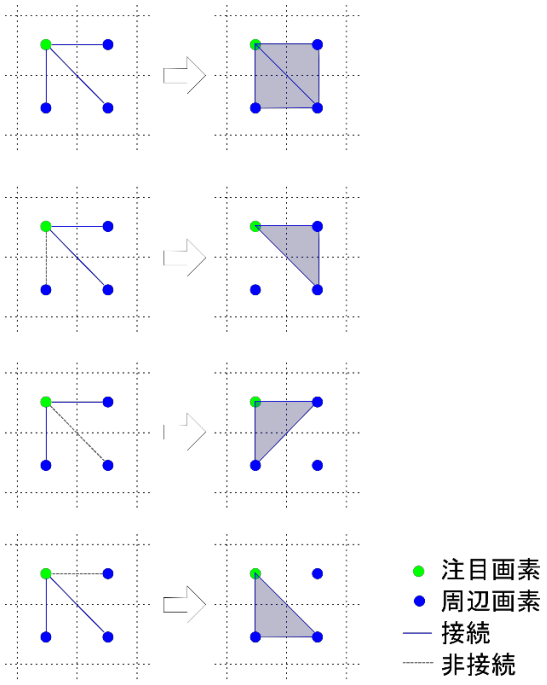

図 11　接続関係とメッシュ化

5.2.2 自由視点映像の生成

生成した3次元形状の各頂点を2次元の光学座標系へ投影し，レンダリングを行う．まず，絶対座標系からカメラ座標系へ座標変換を行う．オドメトリから求めたロボットの基準位置を$(x_s,\ y_s,\ \theta_s)$，基準位置から視点までのオフセットを$(o_x,\ o_y,\ o_z)$とすると，座標変換行列 M は以下のように求まる．

$$\mathrm{M} = \begin{bmatrix} cos\theta_s & sin\theta_s & 0 & -x_t - o_x \\ -sin\theta_s & cos\theta_s & 0 & -y_t - o_y \\ 0 & 0 & 1 & -o_z \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad (14)$$

各点群について，カメラ座標系から 2 次元の光学座標系へ座標変換を行う．光学系への投影には，図 12 に示すような投資投影モデルを用いる．光学座標系での点群の座標を(X, Y, Z)，焦点距離を f，投影点を(x, y)とすると，

$$\mathrm{x} = \mathrm{f}\frac{X}{Z},\ \mathrm{y} = \mathrm{f}\frac{Y}{Z} \quad (15)$$

と求まる．

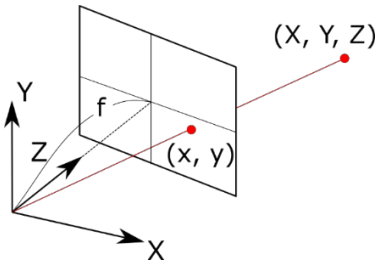

図 12　透視投影モデル

映像生成では，RGB カメラの光学系に透視投影することでレンダリングを行う．カメラキャリブレーションによって推定した RGB-D カメラの外部パラメータから赤外線カメラ座標系と RGB カメラ座標系の間の変換行列を算出し，透視投影により各３次元点群の RGB 画像内での観測位置を求める．予測の基づいた３次元点群の観測位置の移動量に応じて，RGB 画像の画素値を移動させることによって，予測映像を生成する．

5.3 過去フレームの色情報の活用

１フレームごとの映像では，人物の背面など光学的に計測できない領域が存在する．運動予測を適用した結果，映像に欠損が生じてしまうことがある．そこで，過去にその領域が映っていたフレームの情報を利用し，欠損の補償を行う．人物の色情報は，4.3 節の簡易メッシュモデルのテクスチャにより時系列間で保持する．まずメッシュモデルを UV 展開し，２次元のテクスチャ画像との座標の対応付けを行う．撮影された人物領域の各点群について，図 13 のように最近傍のメッシュに垂線を下ろし，交点を計算する．交点がメッシュの内側にある場合，交点の UV 座標を計算し，テクスチャの相当画素の色情報を更新する．映像生成時にテクスチャを適用した簡易メッシュモデルを描写することで，欠損のない人物映像が生成できる．

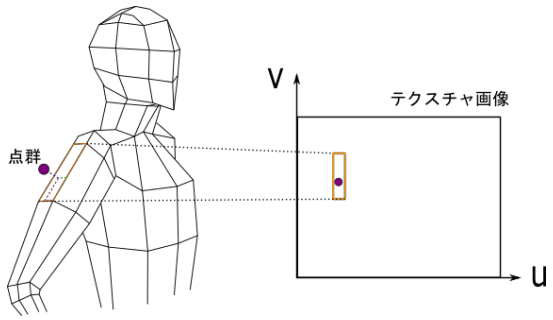

図 13 点群とテクスチャの対応付け

6. 検証実験

6.1 運動予測

実験では，CPU に Intel® Core™ i7-9700 3.0GHz，GPU に NVIDIA GeForce RTX 2080 FH 8GB，RAM: 64GB DDR4，OS に Ubuntu 18.04 を搭載した計算機を使用して処理を実施する．実装言語には Python 3.7 を使用し，TensorFlow [17]を用いて 4 節で紹介したニューラルネットワークを実装する．予測先の時間は，操作遅延と映像の伝送遅延の合計時間として，要望のあった 500ms を対象とする．損失関数は二乗誤差，勾配法には Adam モデルを使用し，ミニバッチサイズ 64 にて 1,000,000 回学習を行う．

学習データには CMU のモーションキャプチャデータ[18]を使用する．CMU のデータセットは歩行，ジャンプ，ドリブル，階段の上り下り等 2,212 種類のモーションデータが収録されている．データ形式には Bounding Volume Hierarchy(bvh)形式が使用されており，ルートである腰は位置について３自由度，54 本のリンクは姿勢についてオイラー角３自由度の表現で記録されている．ここでは，骨格形状で使用する 16 点について，腰の位置を記述している座標系に合わせた３次元座標データを取り出した後，4.1 節で紹介した手法で 31 次元の時系列データを取り出した．収録ファイル単位で４分割し，クロスバリデーションを行う．

評価には直交座標系での各関節位置の誤差平均と，関節角度の誤差平均を用いる．等角加速度モデルによる予測でも同様の評価を行い，本手法の有効性を検証した結果を表３に示す．

表３　評価結果

	等角加速度	提案手法
位置の誤差平均(mm)	76.8	33.73
角度の誤差平均(rad)	0.3505	0.17206

等角加速度モデルと比べ，関節位置，角度それぞれの誤差について半分以下を達成しており，回転系での３次元運動予測が実現されている．

6.2 映像生成

操作者側の処理には 6.1 節と同様のワークステーションを，操作対象には TurtleBot3 を使用して処理を実施した．RGB-D カメラとして Intel RealSense D435 を TurtleBot3 に搭載し，NUITrack [19]により骨格推定と人物領域の抽出を行った．NUITrack では映像内の各人物について，骨格 16 点の３次元座標を取得できる．また，各画素に背景・人物領域のラベル情報，距離情報，位置情報を取得できる．ワークステーション，TurtleBot3 それぞれに ROS 環境を用意し，遠隔操作及び映像の送受信には ROS の Topic 機能を利用した TCP/IP 通信を使用した．原画像を図 14，出力画像を図 15，実際の 500ms 後の画像を図 16 に示す．左足に注目すると，平行移動ではなく関節の角度の変化が発生しており，骨格の挙動予測に合わせた予測映像が生成されていることが確認できる．

7. おわりに

本研究では，RGB-D 映像から推定した人物の３次元骨格形状の一定時間での運動を深層学習を用いて予測し，その予測結果に従って RGB-D 画像から生成される３次元点群を移動させること，またオドメトリ計算によって視点の移動を計算し，自由視点映像のレンダリングを行うことによって，予測映像を生成する手法を提案した．本手法を応用することにより，遠隔操作映像の伝送遅延補償の実現が可能となる．今後は，映像生成においても深層学習を用いることで映像の品質の改善に取り組む．また処理速度を改善しリアルタイム処理を実現させ，実利用における有効性の検証を行う．本研究は，科研費(17H01772)および(19H00806)

の助成を受けたものである.

参考文献

[1] 竹下佳佑，渡邉孝一，佐藤克成，南澤孝太，舘暲: “テレイグジスタンスの研究(第 63 報)-TELESAR3 において許容される通信遅延の検討-”，第 15 回日本バーチャルリアリティ学会大会論文集，2010.
[2] Erwin Wu，Hideki Koike，“FuturePose - Mixed Reality Martial Arts Training Using Real-Time 3D Human Pose Forecasting With a RGB Camera”，2019 IEEE Winter Conference on Applications of Computer Vision (WACV)，Jan．2019.
[3] 大石圭，森尚平，斎藤英雄，"魚眼力メラと 3D-LIDAR を用いた自由視点画像による See-Through Vision"，映像情報メディア学会誌，2017，71.11: J276-J279.
[4] 北原格，大田友一，"多視点映像の融合によるスポーツシーンの自由視点映像生成: 3 次元形状表現用平面の適応的配置"，電子情報通信学会技術研究報告，PRMU，パターン認識・メディア理解，2001，100.633: 23-30.
[5] Zhe Cao，Gines Hidalgo，Tomas Simon，Shih-En Wei，Yaser Sheikh，"OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields"，arXiv preprint arXiv:1812.08008，2018.
[6] Julieta Martinez，et.al，“A simple yet effective baseline for 3d human pose estimation”，ICCV2017，2017
[7] Iro Laina，Christian Rupprecht，Vasileios Belagiannis，Federico Tombari，Nassir Navab．"Deeper depth prediction with fully convolutional residual networks"，2016 Fourth international conference on 3D vision (3DV)．IEEE，p.239-248，2016
[8] 前田雄介; 原崇之; 新井民夫．“躍度最小モデルを用いた動作予測に基づく人間-ロボット協調作業“．日本機械学会論文集 C 編，2002
[9] Jian Bo Yang，Minh Nhut Nguyen，Phyo Phyo San，Xiao Li Li，Shonali Krishnaswany，"Deep Convolutional Newral Networks On Multi Channel Time Series For Human Activity Recognition"，Twenty-Fourth International Joint Conference on Artificial Intelligence，2015.
[10] Ryo Yonetani，Kris Kitani，Yoichi Sato: “Ego-Surfing: Person Localization in First-Person Videos Using Ego-Motion Signatures”，IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)，Vol.40，Issue 11，pp.2749-2761，2018.
[11] Schonberger，Johannes L.，and Jan-Michael Frahm．"Structure-from-motion revisited." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition．2016.
[12] 米本恭三，石神重信，and 近藤徹．"関節可動域表示ならびに測定法." *リハビリテーション医学*32.4 ，1995，207-217.
[13] Ronneberger，Olaf，Philipp Fischer，and Thomas Brox．"U-net: Convolutional networks for biomedical image segmentation." International Conference on Medical image computing and computer-assisted intervention．Springer，Cham，2015.
[14] G.Huang，Z.Liu，L.van der Maaten，K.Q.Weinberger．“Densely Connected Convolutional Networks”，IEEE Conference on Pattern Recognition and Computer Vision (CVPR)，2016.
[15] He，Kaiming，et al．"Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition．2016.
[16] 前山祥一，大矢晃久，and 油田信一．"移動ロボットの屋外ナビゲーションのためのオドメトリとジャイロのセンサ融合によるデッドレコニング・システム." 日本ロボット学会誌 15.8 ，1997，1180-1187.
[17] Abadi, Martín, et al. "TensorFlow: Large-scale machine learning on heterogeneoussystems" 2015. Software available from tensorflow.org.
[18]Hodgins，Jessica．"CMU graphics lab motion capture database."，http://mocap.cs.cmu.edu，2015.
[19] 3DIVI Inc.: “Nuitrack．Body Tracking Software”，2018

図 14　入力画像（t=0ms）

図 15　出力された予測画像（t=500ms）

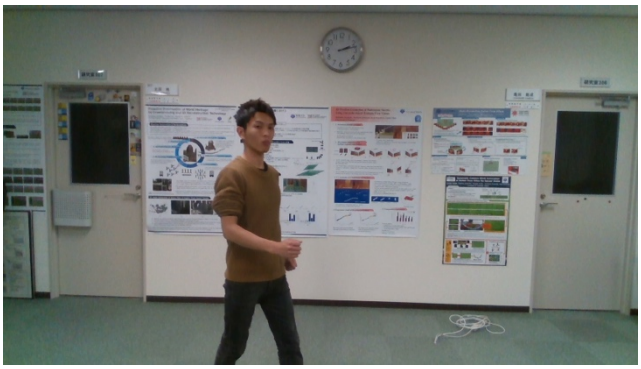

図 16　実際に 500ms 後に撮影した画像