深層学習を用いたテレプレゼンス映像の伝送遅延補償法の検討

上田　樹[*1]　　宍戸　英彦[*1]　　亀田　能成[*1]　　北原　格[*1]

Itsuki Ueda[*1], Hidehiko Shishido[*1], Yoshinari Kameda[*1] and Itaru Kitahara[*1]

あらまし　本稿では，テレプレゼンス映像における伝送遅延の補償を目的とした人物の見え方の予測手法を提案する．RGBD カメラで撮影した映像情報から推定した人物の3次元骨格情報を手がかりに，深層学習によって一定時間が経過した骨格形状を予測する．骨格形状から人物領域形状を表す3次元点群を生成し，その形状情報を用いて予測3次元映像を生成する（人物の見え方変化を予測する）．

キーワード：深層学習，テレプレゼンス，骨格検出，3次元点群，運動予測

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．伝送遅延は遠隔操作の課題の一つである．大きな遅延時間は，操作性の低下だけでなく，誤操作によって危険な状況を招く恐れがある．特に，操作機器の周囲に人物が存在する環境では，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安全な遠隔操作技術が求められている．

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[1]．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[2][3]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，一定時間を映像の遅延時間とすることで，遅延補償の実現を目的とした研究について述べる．

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が必要である．本研究で撮影で対象とする撮影空間は，人物と遠隔操作する機器が混在する環境であるため，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[4]は単眼画像から2次元の骨格を高速に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[5]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[6]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．RGBD 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 RGBD 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，RGBD 画像から3次元人物骨格情報を推定する．

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[7]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 [1]．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[8]．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．

*1 筑波大学

本手法では，RGBD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

竹下ら[9]は，テレイグジスタンスにおいて許容される遅延誤差を検証している．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．

米谷らの研究[10]では，映像中における2次元骨格情報の時系列データを用いて，畳み込みニューラルネットワークによる重心の移動を予測した．利用した2次元骨格情報の時系列データは，画像内での人物の大きさ，カメラ位置姿勢を含めている．遠隔操作のように物体を動かす場合，危険な接触を避けるために奥行情報は重要な意味を持つ．本研究では深層学習に基づき，入出力の骨格情報を2次元から3次元へ拡張する．さらに骨格の重心推定から全ての骨格の位置推定へ拡張する．

3 深層学習を用いた伝送遅延補償法

提案システムのフローチャートを図 1 に示す．RGBDカメラで人物を撮影し，奥行き画像から人物の3次元骨格情報の推定を行う．あわせて，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，RGBD映像から人物領域の3次元点群を生成する．予測した3次元骨格の動きに合わせて3次元点群を移動・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．

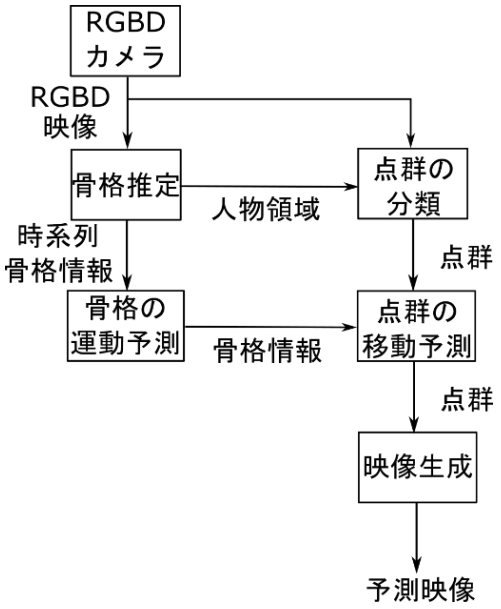

図 1 深層学習を用いた伝送遅延補償法の概要

4 骨格の運動予測

本節では，骨格の時系列データから，未来の骨格を推定する手法について述べる．

4.1 重心位置と進行方向による関節位置の正規化

本方式で入力とする骨格モデルの各関節の3次元座標には，全身の重心位置や運動と，各部位の位置や運動が混在している．そのような状態で深層学習を適用した場合，支配的な（この場合だと全身の）運動の影響を強く受けた予測器が生成されてしまう．そこで，全身の位置と運動を求め，その情報を用いて正規化（全身運動をキャンセル）した関節の位置情報を用いて深層学習を行う．本方式では，腰関節の3次元座標を全身の重心位置とし，その移動全身の運動（進行方向）とする．具体的には，図 2 に示すように，現フレームの腰の位置を地上面へ投影した位置を原点，進行方向を x 軸とするような座標変換を行う．

時系列データの先頭フレームと最終フレームの腰関節の座標を$(\mathrm{x}_0, \mathrm{y}_0, \mathrm{z}_0)$，$(\mathrm{x}_\mathrm{n}, \mathrm{y}_\mathrm{n}, \mathrm{z}_\mathrm{n})$とすると，y 軸回りの回転角度θは$\mathrm{atan2}(\mathrm{z}_\mathrm{n} - z_0, x_n - x_0)$で求められる．座標変換$\mathrm{M}_\mathrm{l}$，絶対座標系への逆変換$\mathrm{M}_\mathrm{g}$は

$$\mathrm{M}_\mathrm{l} = \begin{bmatrix} cos\theta & 0 & -sin\theta & -x_n\, cos\theta + z_n sin\theta \\ 0 & 1 & 0 & 0 \\ sin\theta & 0 & cos\theta & -x_n sin\theta - z_n cos\theta \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$\mathrm{M}_\mathrm{g} = \begin{bmatrix} cos\theta & 0 & sin\theta & x_n \\ 0 & 1 & 0 & 0 \\ -sin\theta & 0 & cos\theta & z_n \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

と表される．深層学習への入力データは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RGBDカメラ座標系に変換する．

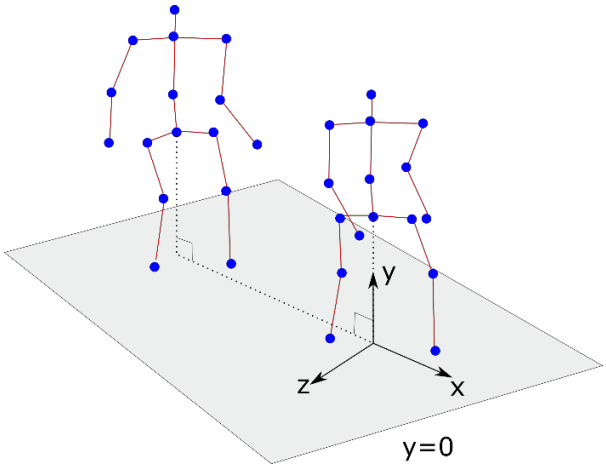

図2　重心位置と進行方向による関節位置の正規化

4.2 深層学習による予測器の生成

各骨格位置の時系列データを入力とし，一定時間経過後の各骨格位置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．

各骨格は16点のモデルを用い，それぞれの直交座標系での3次元座標を使用した48次元で表現する．また入力する時系列データは30fpsで64フレームを使用する．表1にネットワーク構成を示す．Coは1D-Convolution，BNはBatch Normalization，MPはMax Pooling，FCはFull Connectionを表す．

表1　構成したネットワーク

Layer Type	Channel	Kernel Size	Output Size
Input	-	-	48×64
Co+BN+ReLU	64	3	64×62
Co+BN+ReLU+MP	128	3	128×30
Co+BN+ReLU+MP	256	3	256×14
Co+BN+ReLU+MP	512	3	512×6
Layer Type	Input Size		Output Size
FC+ReLU	3072		1024
FC+ReLU	1024		512
FC+ReLU	512		256
FC+ReLU	256		128
FC+ReLU	128		64
Output(Linear)	64		48

ニューラルネットワークの学習には，CMUのモーションキャプチャデータ[11]を使用する．CMUのデータセットは，歩行，ジャンプ，ドリブル，階段等2212種類のモーションデータが収録されている．データ形式は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あわせた3次元座標データを取り出した．

5 人物領域（3次元点群）の移動

本節では，3次元骨格の運動予測結果から，人物の3次元点群を移動させる処理について述べる．

5.1 リンクの姿勢の計算

骨格の関節を結んだリンクから離れた3次元点群を移動するためには，骨格の位置だけでなくリンクの姿勢変化の情報が必要である．しかし16点の関節で構成されるポーズモデルでは各関節の回転軸を一意に定めることができない．また直交座標系での推論を行なっているため，各関節の位置関係に拘束条件はなく，リンクの長さが固定されていない．本方式では，現在の映像の骨格を基準に，各骨格の座標変換行列をフォワードキネマティクスにより計算することにより，この問題を解消する．

図3上段中の実線で示すリンクを点線で示すリンクに合わせる場合を例に説明する．まず平行移動$\mathrm{trans}(q_1 - p_1)$によってリンクの根本を一致させる．リンク$l_1$の動作は，図3中段のような$\mathbf{n_1}$を軸とした角度$\theta_1$の回転である．$\mathbf{n_1}$はリンク方向の外積，$\theta_1$は正であることから内積を用いて回転を求める．

$$\boldsymbol{n_1} = \frac{(\mathbf{p_2'} - \mathbf{p_1}) \times (\boldsymbol{q_2} - \boldsymbol{p_1})}{|(\mathbf{p_2'} - \mathbf{p_1}) \times (\boldsymbol{q_2} - \boldsymbol{p_1})|}$$

$$\theta_1 = \cos^{-1} \frac{(\mathbf{p_2'} - \mathbf{q_1}) \cdot (\boldsymbol{q_2} - \boldsymbol{q_1})}{|(\mathbf{p_2'} - \mathbf{q_1})|\,|(\boldsymbol{q_2} - \boldsymbol{q_1})|}$$

以上の処理により，リンクl_1の座標変換T_1は以下のように求まる．

$$T_1 = \mathrm{trans}(q_1 - p_1)\mathit{trans}(p_1)\mathit{rot}_{n_1}(\theta_1)\mathit{trans}(-p_1)$$

リンク全体にT_1を適用すると，図3下段のような変形になる．リンク2の動きは，リンクの向きを$(\mathbf{q_3} - \mathbf{p_2''})$へ向けることで$\mathbf{p_3''}$の誤差を最小化できる．リンク$l_2$の回転軸$\mathbf{n_2}$，角度$\theta_2$はリンク$l_1$と同様に，以下のように求められる．

$$\boldsymbol{n_2} = \frac{(\mathbf{p}_3'' - \mathbf{p}_2'') \times (\boldsymbol{q_3} - \mathbf{p}_2'')}{|(\mathbf{p}_3'' - \mathbf{p}_2'') \times (\boldsymbol{q_3} - \mathbf{p}_2'')|}$$

$$\theta_1 = \cos^{-1} \frac{(\mathbf{p}_3'' - \mathbf{p}_2'') \cdot (\boldsymbol{q_3} - \mathbf{p}_2'')}{|(\mathbf{p}_3'' - \mathbf{p}_2'')||(\boldsymbol{q_3} - \mathbf{p}_2'')|}$$

リンクl_2までの座標変換T_2は以下のように求まる.

$$\mathrm{T}_2 = \mathrm{T}_1 trans(p_2) rot_{n_2}(\theta_2) trans(-p_2)$$

一般化すると，リンクl_Nまでの座標変換T_Nは以下のように求まる.

$$\mathrm{T}_0 = \mathrm{trans}(\mathrm{q}_0 - \mathrm{p}_0)$$

$$\widehat{\boldsymbol{p_N}} = T_{N-1}\,\boldsymbol{p_N},\ \widehat{\boldsymbol{p_{N+1}}} = T_{N-1}\,\boldsymbol{p_{N+1}}$$

$$\boldsymbol{n_N} = \frac{(\widehat{\boldsymbol{p_{N+1}}} - \widehat{\boldsymbol{p_N}}) \times (\boldsymbol{q_{N+1}} - \widehat{\boldsymbol{p_N}})}{|(\widehat{\boldsymbol{p_{N+1}}} - \widehat{\boldsymbol{p_N}}) \times (\boldsymbol{q_{N+1}} - \widehat{\boldsymbol{p_N}}')|}$$

$$\theta_N = \cos^{-1} \frac{(\widehat{\boldsymbol{p_{N+1}}} - \widehat{\boldsymbol{p_N}}) \cdot (\boldsymbol{q_{N+1}} - \widehat{\boldsymbol{p_N}})}{|(\widehat{\boldsymbol{p_{N+1}}} - \widehat{\boldsymbol{p_N}})||(\boldsymbol{q_{N+1}} - \widehat{\boldsymbol{p_N}}')|}$$

$$\mathrm{T}_N = \mathrm{T}_{N-1} trans(p_N) rot_{n_N}(\theta_N) trans(-p_N)$$

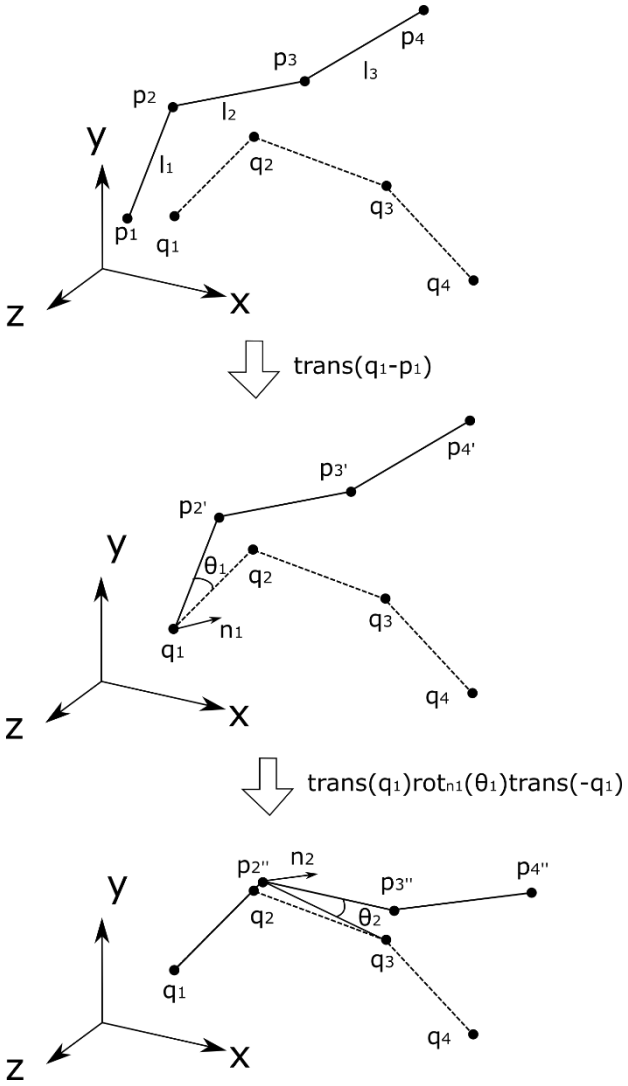

図 3　リンクの姿勢計算

5.2 点群の移動の予測

図 4 に示すように，人物領域の各3次元点群を最近傍のリンクと結び付け，対応するリンクの座標変換を適用する. リンクの結び付けでは，各リンクを線分とした時の各点群との距離によって評価し，最小のものと結びつける. 関節の外側では距離の等しい点群が存在する. そのように点群がリンクの線分に垂線を取れない場合には，距離のリンク方向成分も記録し，リンク方向成分の小さいものを選択する.

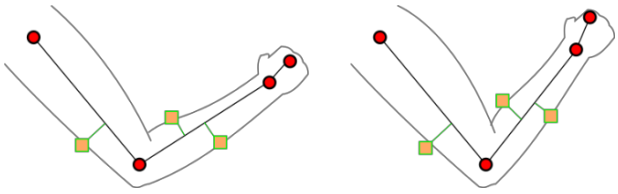

図 4　点群の移動予測

6 映像生成

本節では，映像生成の手法について述べる．映像生成には，図5に示す透視投影モデルを用いる．まず，各点群について，絶対座標系から撮影対象の光学座標系へ座標変換を行う．光学座標系での点群の座標(X, Y, Z)，焦点距離 f，投影点(x, y)とすると，

$$x = f\frac{X}{Z},\ y = f\frac{Y}{Z}$$

と求まる．

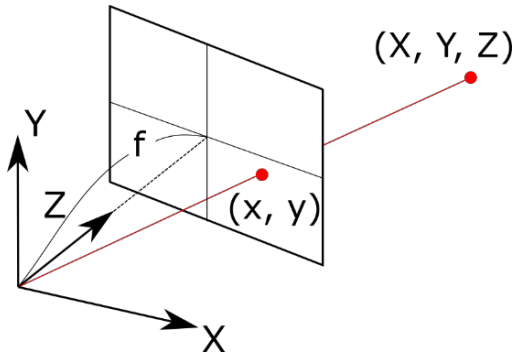

図5 透視投影モデル

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 RGBD 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．

7 提案手法の実証実験

実験では，CPU: Intel(R) Core™ i7-7700HQ 2.80GHZ, GPU: NVIDIA Geforce GTX 1060, メモリ: 16.0GB RAM を装備したノート PC を使用して処理を実施する．開発環境は Visual Studio 2019 を用い，言語には C++と CUDA を使用する．入力する RGBD カメラには Intel RealSense D435 を使用し，NUITrack[12]により骨格推定と人物領域の抽出を行う．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，色情報を取得できる．ラベルごとに各画素の距離，色から点群を生成することで，分類済みの点群を出力する．

骨格の運動予測は,4節で紹介したニューラルネットワークを使用する．深層学習は Tensorflow を用いて実装し，学習済みのパラメータをバイナリ形式で書き出した後，CUDA で実装した推論機から読み込み使用する．損失関数は2乗誤差，勾配法には AdaGrad モデルを使用し，バッチサイズ 32 にて 1000000 回学習を行う．

点群の移動予測は，NUITrack が出力した現在の骨格を基準に，ニューラルネットワークが出力した骨格への座標変換を計算する．RGBD 映像から取得した点群を現在の骨格の各リンクに結び付け，対応する座標変換によって移動させる．

移動予測を適用した点群を，RealSense の RGB カメラへ透視投影モデルで投影することで映像生成を行う．RGBD カメラのカメラパラメータは，チェスボードの撮影結果を用意し，OpenCV のキャリブレーション API を利用して設定した．キャリブレーションにはチェスボードを写した複数の画像を撮影し，cv::CalibrateCamera によって各カメラのパラメータを，cv::stereoCalibrate によって赤外線カメラと RGB カメラ間の変換行列を取得した．

8 予測映像の生成

RealSense の映像から，時間補償を行った結果を示す．原画像を図 6 に示す．骨格の挙動予測結果を図 7 に示す．出力画像を図 8 に示す．骨格の挙動予測では，平行移動ではなく関節の角度の変化まで予測されている．また，出力画像では，骨格の移動に合わせて色が移動していることが分かる．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．

図6 原画像

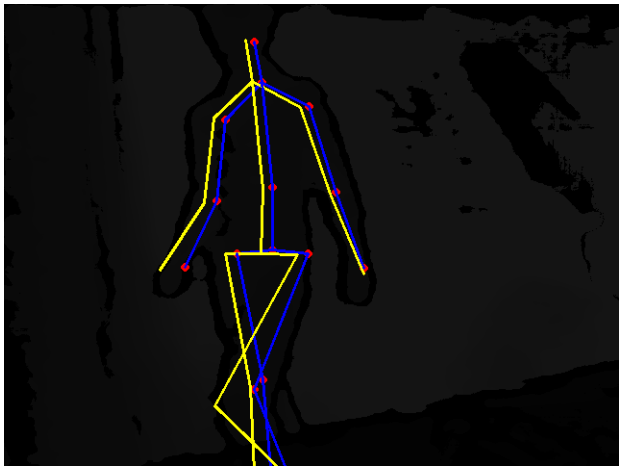

図 7　骨格の挙動予測（青：入力された骨格，黄：推定された骨格）

図 8　出力画像

9　おわりに

本研究では，RGBD 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移動を深層学習を用いて予測し，その予測結果に従って RGBD 画像から生成される3次元点群を移動させる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，テレプレゼンス映像の伝送遅延補償の実現が可能となる．今後は，挙動予測に適切な拘束条件を追加することで，予測の安定化に取り組む．

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]　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

[2]　大石圭，森尚平，斎藤英雄，"魚眼カメラと 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.

[3]　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.

[4]　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.

[5]　Julieta Martinez, et.al, A simple yet effective baseline for 3d human pose estimation, ICCV2017

[6]　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

[7]　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 (機械力学，計測，自動制御)．*日本機械学会論文集 C 編*，2002

[8]　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.

[9]　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.

[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]　Carnegie Mellon University: "CMU Graphics Lab Motion Capture Database", http://mocap.cs.cmu.edu/

[12]　3DIVI Inc.: Nuitrack．Body Tracking Software．3DIVI Inc. 2018