空間再構成のための
遮蔽に頑健な骨格推定技術に関する一検討

岡見　和樹[*1]　松村　誠明[*1]　能登　肇[*1]　木全　英明[*1]

Abstract --- 近年，スタジアムなどの ICT 化が進行する中で，多視点映像や 360 度映像の配信など，単なる映像視聴の枠を超えたサービスの検討が盛んに行われている．我々は将来的な映像視聴サービスとして自由視点映像に着目し，フィールド内に立ったような近接した視点からの高臨場な映像生成を目指している．しかし，近接した視点からの高品質な映像生成は，既存の空間再構成方式では実現困難である．そこで，映像内の人物の骨格推定を行い，事前に作成しておいた当該選手の高品質な CG モデルに割り当てる空間の再構成手法を提案する．本稿では，従来技術のようにカラー画像のみを用いて骨格を直接推定するのではなく，カラー画像から推定対象のシルエット画像を推定し，カラー画像と併用して骨格推定をする．これにより，被写体が遮蔽物に隠れた環境下において，従来手法に対する一定の優位性が確認できたため報告する．

Keywords: 自由視点映像，空間再構成，骨格推定，DNN

1 はじめに

近年，スタジアムなどの ICT 化が進行し，多視点映像や 360 度映像の配信など単なる映像視聴の枠を超えたサービスの検討が盛んに行われている．我々は，将来の映像配信サービスとして，競技場内で視点位置を任意に変更できる自由視点映像と呼ばれる映像視聴に注目している．自由視点映像は，競技場内を自由に歩き回ったり，特定の選手と同じ目線に立ったり，更にはボール目線での映像視聴を行うといった，カメラで撮影することが困難な映像を提供することが可能であり，現実を超えた超高臨場な映像視聴体験を実現できる可能性を秘めている．

自由視点映像を実現する手法には様々なものが存在しており，有名な手法の一つとして，ビルボードを用いる手法[1]が挙げられる．この手法は，ビルボードと呼ばれる 3 次元平面に，画像中の被写体領域を投影することで，高速に自由視点映像を合成可能であるが，被写体の細かい凹凸の再現や視差表現は行えないため，フィールド内部などの選手に近い位置からの視聴を実現することには不向きである．被写体の 3 次元形状を復元する手法[2]も提案されているが，これらの手法の多くは，大量のカメラやグリーンバックなど予め調整された環境を用いて被写体の 3 次元形状を厳密に再現している．そのため，比較的小規模な空間を復元するときに有効であるが，撮影環境の制約が大きく，スタジアムなど大規模な空間での運用が難しい．また，環境の制約を緩和可能な深度センサを用いた手法[3]も提案されているが，一般に深度センサは解像度が低く，データを取得可能なレンジも狭いため，こちらもスタジアムなどの比較的大規模な空間での運用は困難である．また，これらに共通の問題として，撮影されたカラー画像の解像度が低い場合に被写体のテクスチャがぼけてしまうことや，遮蔽物に被写体が隠れた場合に合成される 3 次元形状に欠損が生じることなどが挙げられる．特にチーム対抗で行うスポーツは，一定の領域で多くの選手が入り乱れることが頻発するため，自由視点映像配信サービスの実現に向けて，遮蔽物に対する安定性の向上は解決すべき課題の一つである．

そこで我々は，競技場で撮影した実際の映像から選手の動作を推定し，その推定結果を事前に作成した当該選手の高品質な CG モデルの動作として割り当てることで，上記の問題を解決することを提案する．推定した動作情報に基づいて高品質モデルを変形させる手順を踏むことで，前述した遮蔽物による欠損やテクスチャの低解像化といった問題を解決する．背景などを事前に CG で作成した上で，上記の処理空間内に存在する全選手に対して行うことで，CG として空間全てを再構成する．近年の映画やゲームコンテンツなどで見られるように，CG の品質は非常に高い水準に達しているため，本手法を用いることで，現実と見紛うような品質の自由視点映像を生成することが可能となる．提案する空間再構成の手順を図 1 に示す．

*1 NTT メディアインテリジェンス研究所

図 1　空間再構成手法

Fig.1　Our field reconstruction method

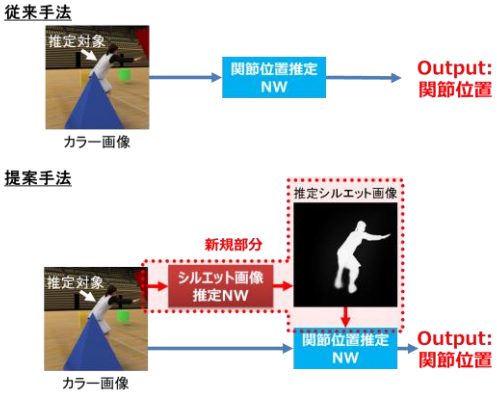

図 2　提案する骨格推定手法

Fig.2　Proposed pose estimation method

上記の手段による空間再構成の実現には，選手の骨格情報を用いる．しかし，選手の動きを妨げるモーションセンサーの装着は不可能であるため，映像から動作を推定するための骨格推定技術が必要となる．以降では，映像から骨格推定を行う技術に関連する既存手法とその課題について触れる．

2　関連研究

映像からの骨格推定は，古くは人体の構造を仮定したり，最適な特徴量をシーンに合わせて規定して機械学習を行ったりすることで，腕や肘といった各パーツの位置を推定していた．しかし，近年の畳み込みニューラルネットワーク(CNN)をベースとした深層学習(Deep Learning)の発展により，それら要素の自動決定と最適化が進み，より頑健な推定が可能となった．Deep Learning を用いた骨格推定技術の中で特に有名な物として OpenPose[4]と呼ばれる手法が挙げられる．この手法は，各関節点のマップ推定と同時に，Part Affinity Fields(PAFs)と呼ばれる各関節の関連性も同時に推定する．これにより，映像内に複数人物が存在していても，人物間の誤認を抑えた高精度な骨格推定が可能となっている．

しかし，映像内の被写体が遮蔽物に隠れている場合，推定精度が低下したり，関節を検出できなくなったりといった問題を抱えている．前章でも触れた通り，スタジアムでの運用及び推定した情報を事前作成した CG モデルに割り当てることを考慮すると，遮蔽物が存在する場合においても検出率を高く保つことが課題であるといえる．

3　提案手法

本稿では，上記の課題を解決するために，従来技術と同様にカラー画像のみを用いて関節位置を直接推定するのではなく，まず，カラー画像から推定対象のシルエット画像を推定し，その後，カラー画像と推定したシルエット画像を併用して関節推定器のインプットとすることで骨格推定を行う．導入したシルエット画像が骨格位置のガイドの役割をこなすことで，遮蔽物が存在する場合であっても頑健な推定が可能となると期待できる．従来技術と本手法の構成について図 2 に示す．

4　データセット作成及びネットワーク構成

4.1　データセット作成

本手法では，被写体が遮蔽されている部分も含めて関節の検出を行うことを目的とし，関節位置推定の際にシルエット画像を併用する．そのため，用意するシルエット画像は画像内で見えている範囲のみをセグメンテーションしたものではなく，遮蔽されている領域も含めてセグメンテーションしたものが望ましい．公開されているデータセットに画像を合成して遮蔽を作り出している手法[5]や，元々遮蔽が存在している画像から推測して人手でデータセットを作成する手法[6]などが提案されているが，本手法では CG を用いてデータセットを作成する．CG を用いてシーンを作成し各種情報を出力することで，映像に不連続な改変を加えることなく，正確なシルエット画像や関節位置情報を出力することが出来るためより効果的なデータセット作成が可能になると考えられる．

今回はモーションキャプチャスーツを用いて取得したバスケットボールの選手の動作およびモデルを CG で作成し，注目選手を定め，その選手を追跡するよう設定した複数のカメラから各種データを出力することとする．

表 1　作成したシーンの内容

Table 1　Detail of CG scene

項目	詳細
概要	バスケットのディフェンスモーション
長さ	597Frame
カメラ	10 台
ソフトウェア	3dsMax 2017，V-ray 3.40.03

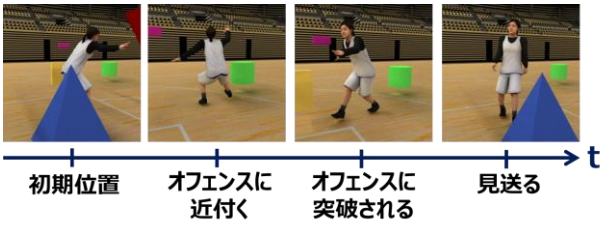

図 3　CG シーンの例

Fig.3　Examples of CG scene

なお，遮蔽物は本来ボールや他選手が望ましいが、本試行では基礎データ取得を目的とし，CG で作成したプリミティブを選手周辺に配置することで代用した．表 1 に作成したシーンの詳細を，図 3 にシーンの一例を記す．

4.2 シルエット画像推定ネットワーク

図 2 の下部に示すように，まずカラー画像から遮蔽されている領域も含めてシルエット画像を推定する．画像からのセグメンテーションには CNN をベースとした手法が多く採用されており，本手法も CNN をベースとしたシルエット推定を行う．構築したネットワークを図 4 に示す．このネットワークは画像から深度画像を推定するために用いられるネットワーク[7]を参考にしており，中間出力として深度画像を出力するように学習を行う．深度画像を推定することで，画像内の相対的な構造を捉えることが可能となり，より正確なシルエットが出力されることが期待できる．

4.3 関節位置推定ネットワーク

カラー画像と，前節に記載したシルエット画像推定ネットワークから推定されたシルエット画像を用いて，画像上の関節位置を推定する．この時推定される関節の位置と個数を図 5 に示す．関節位置を推定するために用いるネットワークは高精度かつ使用に制限がない ArtTrack[8] をベースに，RGB のカラーチャンネルに加えてアルファチャンネル部分にシルエット画像を追加できるよう改変したものを用いる．

5 実験

5.1 実験環境

実験に使用したフレームワークの詳細を表 2 に示す．

5.2 シルエット画像推定ネットワーク学習結果

シーンを撮影しているカメラの中から 5 台のカメラを選択し，全 597Frame のシーンから 3Frame ごとに抽出した，計 995 枚のカラー画像と，深度画像，シルエット画像を学習に使用した．画像サイズはそれぞれ 224x224 である．また，オプティマイザは学習率を 1.0e-5 に設定した Adam，深度画像の loss は mse，シルエット画像の loss は binary_crossentropy，loss の重みは深度画像を 0.25，シルエット画像を 0.75，バッチサイズは 2 とそれぞれ設定して学習を行った．学習の様子を図 6 に示す．シルエット画像の loss はおよそ 250 エポック付近まで減少を続

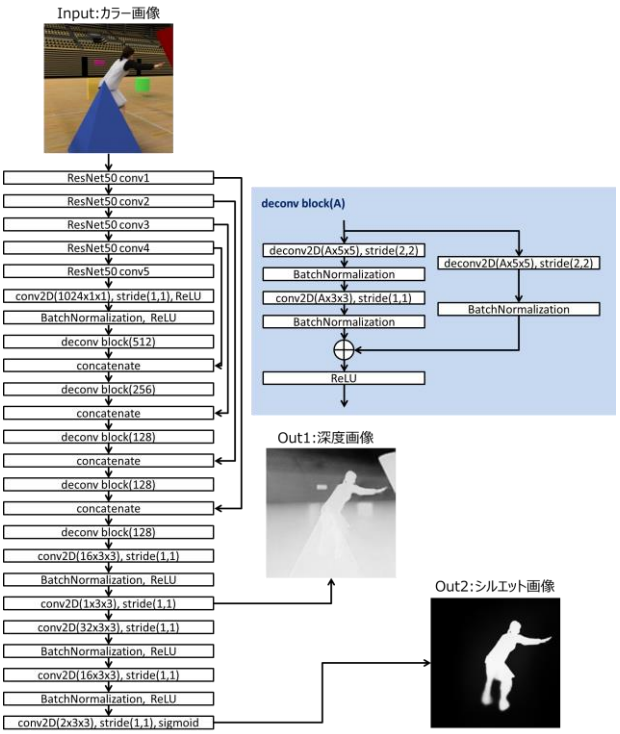

図 4　シルエット画像推定ネットワーク

Fig.4　Silhouette estimation network

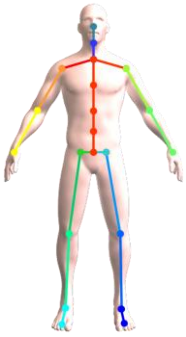

図 5　推定される関節の構成

Fig.5　Estimated joint structure

表 2　実験環境

Table 2　Experiment environment

項目	仕様
CPU	Intel Core i7-6700(3.40GHz)
メモリ	32.0GB
GPU	NVIDIA GeForce GTX1080
フレームワーク	Anaconda 3 python 3.5.2 keras 2.0.4 tensorflow-gpu 1.1.0 CUDA 8.0 (cuDNN 6.0)

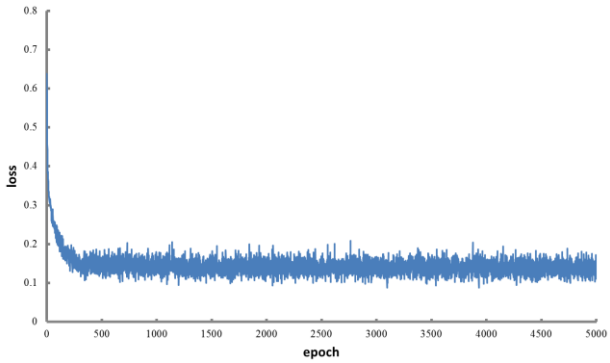

図 6 シルエット画像推定ネットワークの学習過程

Fig.6 Learning process of silhouette estimation network

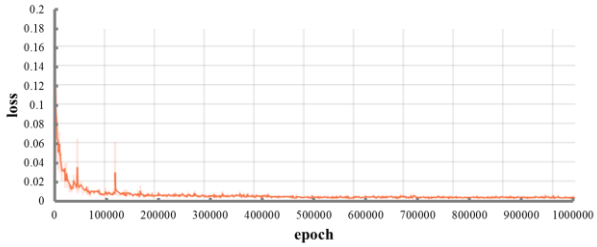

図 7 関節位置推定ネットワークの学習過程

Fig.7 Learning process of joint estimation network

けた後は多少振動をしつつ， 0.15 付近で安定した．

5.3 関節位置推定ネットワーク学習結果

前節で述べたシルエット画像推定ネットワークの学習で用いたものと同一のカメラ 5 台を選択し，全 597Frame のシーンから全ての Frame を用いて，計 2985 枚のカラー画像と，推定したシルエット画像，画像上の関節位置情報をそれぞれ学習に使用した．また学習条件は[8]に準ずるものとした．学習の様子を図 7 に示す．loss はおよそ 30000 エポック付近まで急激に減少した後，緩やかに減少し 100000 エポック以降では，ほぼ 0.01 を下回る数値を示した．

5.4 仮想環境での適用結果

前述した学習結果を用いて，まずカラー画像からシルエットの推定を行う．ここでは，各学習で用いたカメラとは異なるカメラ 5 台を選択し，全 597Frame のシーンから全ての Frame を用いて，計 2985 枚のカラー画像を入力とする．深度画像及びシルエット画像の推定結果の一例を図 8 に示す．左から，入力画像，シルエット画像の真値，深度画像の同時推定なしのシルエット画像推定結果，深度画像の同時推定ありのシルエット画像推定結果を示し，上二列は被写体が遮蔽されている場合の結果，下二列は被写体が遮蔽されていない場合の結果をそれぞれ示す．また，推定結果の精度を表 3 に示す．深度画像の同時推定なしの場合と，深度画像の同時推定ありの場合それぞれに対して，使用した 5 つの各

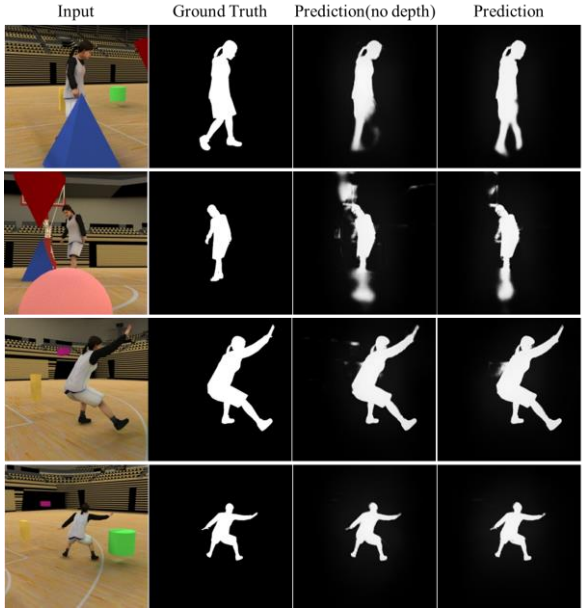

図 8 シルエット画像推定結果

Fig.8 Results of silhouette images estimation

表 3 シルエット画像推定結果精度

Table 3 Accuracy of estimated silhouette images

	Prediction IoU (no depth)	Prediction IoU
Scene A	89.0%	90.2%
Scene B	79.6%	78.0%
Scene C	79.2%	82.4%
Scene D	88.3%	89.0%
Scene E	79.7%	81.4%
Average	83.2%	84.2%

シーン全 Frame の IoU（Intersection over Union）値の平均を記載している．

図 8 から，被写体が遮蔽されていない場合については，深度画像の同時推定の有無に係らず，真値に近い結果が得られていることがわかる．被写体が遮蔽されている場合に関しては，深度画像の同時推定を行った場合，深度画像の同時推定を行わない場合に比べて，遮蔽されている部分の欠損が少なく，また被写体以外の不要な部分の推定が抑制されていることが分かる．

表 3 を見ると深度画像の同時推定を行った場合の IoU 値向上はシーンの全体平均でおよそ 1 ポイント程度に留まっている．しかし，上述したように，被写体が遮蔽されていない場合については，深度画像の同時推定の有無に係らず類似した結果が得られており，被写体が遮蔽されている場合については推定結果が大きく異なることに鑑みると，遮蔽が生じている場合にはより高い効果が見込めると考えられる．例えば，一枚目のインプット画像における推定結果の IoU 値は，深度画像の同

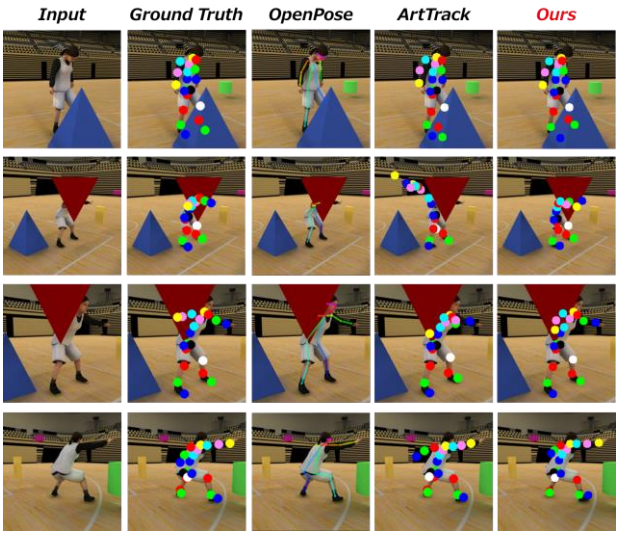

図 9　関節位置推定結果

Fig.9　Results of joints estimation

表 4　関節検出率

Table 4　Joints detection rate

	OpenPose	ArtTrack	Ours
Scene A	59.1%	79.6%	88.1%
Scene B	51.4%	33.3%	59.0%
Scene C	70.5%	24.0%	69.5%
Scene D	63.7%	83.9%	91.0%
Scene E	53.9%	38.6%	51.1%
Average	59.7%	51.9%	71.8%

時推定無の場合 66.2%，同時推定有の場合 69.5%であり，二枚目のインプット画像における推定結果の IoU 値は，深度画像の同時推定無の場合 56.4%，同時推定有の場合 71.3%であり，被写体が遮蔽されている画像に対しては特に顕著な改善が確認されている．

次に，上記で推定したシルエット画像と，その際に使用した同一のカラー画像を用いて画像上の関節位置を推定する．推定結果の一例を図 9 に示す．左から，入力画像，真値，OpenPose[4]による推定結果，ArtTrack[8]による推定結果，本手法による推定結果をそれぞれ示す．また，推定結果の精度を表 4 に示す．ここでは，PCP(Percentage of Correct Parts) 0.5 による検出率を指標とし，使用した 5 つの各シーン全 Frame の平均値を記載する．なお，OpenPose に関しては図 5 の関節と一致するもののみ評価に使用することとした．

図 9 より，Openpose に関しては，遮蔽されていない部分の検出精度は高いものの，遮蔽部分は検出自体が行われていないことが見て取れる．ArtTrack に関しては，検出自体はほぼすべての関節で行われているものの，遮蔽物や背景の色合いに大きく影響を受けていることが分かる．一方，本手法では，遮蔽されていない部分に関しては従来技術と同様であるが，遮蔽されている部分の検出もある程度行われており，シルエットを導入することで，ArtTrack よりも精度が向上していることが見て取れる．また，背景などとの色味の類似による影響も抑制されていることが分かる．

表 4 を見ると，全体として検出率が大きく向上し，平均で 10 ポイント以上向上している．OpenPose との比較に関しては，今回使用したデータセットが限定的であるため，優位性が完全に示せたとは言い難いが，一定の可能性は示せたといえる．ArtTrack との比較に関しては，シルエット推定の効果が表れた結果であるといえる．

6 まとめ

本研究では，カラー画像のみを用いて骨格を直接推定するのではなく，カラー画像から推定対象のシルエット画像を推定し，カラー画像と併用して骨格推定をする手法を提案した．その結果，被写体が遮蔽物に隠れた環境下においても，精度良く画像上の関節位置を推定できることが確認できた．しかし，今回一つのシーンでしか学習を行っていないため，あくまで限定された仮想環境下での優位性が示せたに過ぎないと考えられる．よって，学習データのバリエーション増加，及び他の様々なシーンや実写画像を用いた比較等を行うことで，本手法の汎用性を検証していく必要がある．

参考文献

[1] Y. Ohta, I. Kitahara, Y. Kameda, H. Ishikawa, and T. Koyama, “Live 3D Video in Soccer Stadium,” International Journal of Computer Vision (IJCV), 75(1), pp.173-187, 2007.

[2] S. Nobuhara, W. Ning, and T. Matsuyama, “A real-time view-dependent shape optimization for high quality free-viewpoint rendering of 3D video,” 3DV, 2014.

[3] 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, 34(4), 2015.

[4] Z. Cao, T. Simon, S. E. Wei, and Y. Sheikh, “Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields,” CVPR 2017.

[5] K. Li, and J. Malik, “Amodal instance segmentation,” ECCV 2016.

[6] Y. Zhu, Y. Tian, D. Mexates, and P. Dollar, “Semantic Amodal Segmentation,” arXiv: 1509.01329, 2015.

[7] Y. Kuznietsov, J. Stückler, and B. Leibe, “Semi-Supervised Deep Learning for Monocular Depth Map Prediction,” CVPR 2017.

[8] E. Insafutdinov, M. Andriluka, L. Pishchulin, S. Tang, E. Levinkov, B. Andres, and B. Schiele, “ArtTrack: Articulated Multi-person Tracking in the Wild,” CVPR 2017.