特別研究報告書

CNNによる深度画像を用いた視点不変な人物姿勢推定

指導教員　　美濃　導彦 教授

京都大学工学部情報学科

髙橋　龍平

平成30年2月8日

CNNによる深度画像を用いた
視点不変な人物姿勢推定

髙橋　龍平

内容梗概

街角や店舗内，駅構内などに深度センサー付きカメラを設置し，そこから得られる人物の画像に対し人物姿勢推定を行えば，不審な行動をとる人物の検知に役立てたり，人物の手に取った商品を推定し，客がどのような商品に興味を示したかを把握することによるマーケティングに活用したりできる．人物姿勢推定とは，人物の写っている画像から各関節の3次元空間中での位置を判定する問題のことである．本研究では，これらのアプリケーションへの応用を考慮した，深度画像に対する学習ベースで人物姿勢推定を行う手法を提案する．

人物姿勢推定を行うにあたって，視点変化の問題，自己隠蔽の問題が存在する．視点変化の問題とは，人物に対するカメラの相対的位置によって人物の見え方が変化するというものであり，それをカバーするためには，相対的位置に応じた学習データを用意する必要がある．自己隠蔽の問題とは，人物の頭部や胴体によって腕や下半身が隠れてしまうというものであり，自己隠蔽により見えていない関節の位置を推定することは難しいうえ，自己隠蔽の可能性を考慮せずに推定すると全く違う箇所に関節位置を推定してしまう可能性がある．このことから，画像から関節が見えているかどうかを判断し，見えていない関節の位置を推定しないようにするのが望ましい．

本研究では，モーションキャプチャデータにより様々な姿勢をとらせた人物モデルを用い，正面，上方など様々な方向から深度カメラで撮影した深度画像を用意し，大量のデータを効果的に学習できる畳み込みニューラルネットワーク (CNN) で学習することで視点変化の問題に対処する．また，人物の関節位置の推定を深度画像から直接行うのではなく，人物を31部位に分け，一旦人物の深度画像の各画素が人物のどの部位に属するかを推定し，それに基づいて関節位置を推定する．このように，画素ごとに人物のどの部位が写っているか判定することで，部位推定の時点でカメラに関節のある人物表面が写っているかどうかがわかり，関節に自己隠蔽が発生しているかどうか推定できるようになる．自己隠蔽により画像中に存在していない部位ラベルに対応する関節は関節位置推定を行わないようにし，自己隠蔽の問題に対処する．

CNN を利用する場合は，損失関数を考慮する必要がある．面積の小さいラベルはセマンティックセグメンテーションにおいて欠損が発生しやすい．関節位置に対応する部位ラベルは，後に行う関節位置推定での精度を向上させるため，面積を小さくしている．部位ラベルが欠損してしまうとそのラベルに対応する関節位置推定ができなくなってしまうため，CNN は面積の小さい部位ラベルを欠損してしまうことなく頑健に推定できるように学習しなければならない．そこで，面積の大きい部位ラベルほど小さな重みを与え，面積の小さい部位ラベルほど大きな重みを与えるような重み付けを行う損失関数を提案し，面積の小さな部位ラベルの学習を優先して学習させることで欠損の問題に対処する．

本研究の提案手法の有効性を示すため，部位ラベル推定に特化した提案手法と，セマンティックセグメンテーションのベースラインとして用いられる従来手法で部位ラベル推定精度を比較する実験，視点変化の問題や自己隠蔽の問題にどれほど対処できているか確認する実験，提案手法と部位ラベル推定に基づく人物姿勢推定の従来手法で関節位置推定精度を比較する実験，実画像を適用した場合に提案手法がどれほど正しく推定できるか確認する実験を行う．実験の結果，様々な視点の深度画像を学習させれば視点の変化に対応できることがわかり，さらに視点固定でのみ機能する従来手法よりも高い精度を達成した．一方，実画像に対しては提案手法を適用する前に背景差分を行う必要があるが，その時に発生するノイズなどの影響により精度が低下し，現状では実画像に対しての提案手法の適用は難しいことがわかった．合成画像だけでなく，実画像に対しても頑健な人物姿勢推定ができるような手法を提案することが今後の課題である．

Viewpoint Invariant Human Pose Estimation Using Depth Images By CNN

Ryuhei TAKAHASHI

Abstract

Cameras with depth sensor mounted in streets corners, stores, or stations can be used for estimating human pose. Human pose estimation can be used to detect a person who takes suspicious behavior and for marketing by estimating items taken by customers. Human pose estimation determines the 3D position of each human joint from the image. Human pose estimation by using depth images is robust to change of clothes and lighting, and it also has an advantage that it does not acquire information related to privacy such as face. There are two problems in learning-based human pose estimation. The first problem is a viewpoint change problem, that is, a problem that the appearance of person varies with camera location. Training data have to deal with viewpoint changes, for that reason, the training data should contain images taken from various viewpoints. The second problem is a self-occulusion problem, that is, a problem that arms and lower body are occuluded by the head and upper body. It is difficult to estimate the position of joints that can not be seen, and there is a risk of estimating the joint position at totally different place when estimating without considering of the self-occlusion. From this, it is preferable to determine whether each joint are visible or not from the image.

In this thesis, in order to deal with viewpoint changes, depth images captured by depth cameras from various directions are used as the training data. In addition, instead of directly estimating the joint position, our method divides the person image into body parts. Since a number of them is enormous, they are trained by convolution neural networks which are capable of effectively training a large amount training data. Human is described as 31 body parts, then each pixel is classified to one of the 31 body parts. The center point of pixels that classified to its corresponding body parts is regarded as the joint. By classifying pixels and calculating the number of pixels classified each body part, it is possible to determine whether each joint is observed by the camera. In order to constract the classifier that can estimate which part corresponds to each pixel,

pairs of person depth image and part labeled image are used as training data. A CNN is trained so that it outputs images with correct label when depth image is given as input. This is a kind of semantic segmentation for determining objects for each pixel. The label with small area tends to be defective in semantic segmentation. The part label corresponding to the joint position is reduced in area in order to improve accuracy in later joint position prediction. When the part labels are missing, the position of joint corresponding to that label can not be predicted, so CNN should be trained so as to be robustly estimatable without missing small part labels. Therefore, we propose a loss function that gives smaller weights to larger part labels and larger weights to smaller part labels. By learning small part label with high priority, we address the problem.

In order to demonstrate the effectiveness of our method, several experiments were conducted. (1)Comparing our method that specializes in part label estimation with baseline of semantic segmentation. (2)Confirming the ability to deal with the two problems, viewpoint change and self-occulusion. (3)Comparison of joint position prediction accuracy with conventional method of human pose estimation based on part label estimation. (4)Confirming performance for real images. As a result of experiments, we found that learning depth images of various viewpoints can deal with viewpoint changes, and also achieved higher precision than the conventional method functioning only at fixed viewpoint. On the other hand, experimental results with real images shown that background subtraction is necessary beforehand. Background subtraction decreases the precision due to noise on depth images. There is a problem to apply real data. The future task is to improve the accuracy of the real data.

CNN による深度画像を用いた
視点不変な人物姿勢推定

目次

第1章　序論

街角や店舗内，駅構内などに設置されているカメラから得られる人物の画像に対し人物姿勢推定を行えば，不審な行動をとる人物の検知に役立てたり，手に取った商品を推定することで来客者がどのような商品に購買意欲を示したかを推定し，それをマーケティングに活用したりできる．ここで，人物姿勢推定とは，人物の写っている画像から各関節の3次元空間中での位置を推定する問題である．本研究では，人物姿勢推定を頑健に行う手法を提案する．

人物姿勢推定には，大きくモデルベースの手法と学習ベースの手法が存在する．モデルベースの手法は，画像に写っている人物の部位の大きさや関節間の距離などを推定し，それを元にボーンモデルを生成しフィッティングさせるなどして人物の姿勢を求めるトップダウン的手法である．学習ベースの手法は，様々な姿勢をとった人物の写った画像とその人物の関節位置のデータとのペアをあらかじめ用意し，それらを学習することで各関節の位置を推定できるようにして人物の姿勢を求めるボトムアップ的手法である．モデルベースの手法では，画像1枚のみから人物の部位の大きさなどを全て推定するのは，単純な姿勢でない限り不可能であるという問題点がある．一方，学習ベースの手法はこのような問題は発生せず，近年の機械学習の分野の発展に伴い，モデルベースの手法よりも注目されている．そこで，本研究においても学習ベースの手法を用いる．

監視カメラで撮影された画像から人物姿勢推定を行うにあたって大きく4つの問題がある．1つめの問題は，これは人物姿勢推定に限った話ではないが，監視カメラで不特定多数の人物を撮影することはプライバシーの侵害にあたりうるという問題である．2つめの問題は，人物の服装や照明変化に多くのバリエーションが存在するという問題である．3つめの問題は，人物に対するカメラの相対的位置によって人物の見え方が変化するという視点変化の問題である．カメラと人物間の距離が同じでも，人物を上方から写した画像では，正面から写した画像と比べて，頭が大きく写り，足が小さく写る．すなわち，カメラの向きによって，画像に写る人物の各部位の大きさの比率が異なる．視点変化の問題は，単純に人物の向きを変えるだけで再現することはできず，再現するにはカメラの位置自体を合わせなければならない．4つめの問題は，人物の頭部や胴体によって腕や下半身が隠れてしまう自己隠蔽の問題が生じることである．見え

ていない関節の位置を推定することは難しいうえ，自己隠蔽の可能性を考慮せずに推定すると全く違う箇所に関節位置を推定してしまう可能性がある．例えば，自己隠蔽により右手が全く見えていない状態で右手の位置を推定しようとした時，画像中に右手が存在しないため，右手に形状が似ている左手を誤って右手と推定してしまう，というようなケースが考えられる．このことから，画像から関節が見えているかどうか判断し，見えていない場合はその関節には関節位置推定をしないようにするのが望ましい．前述した監視カメラは，他の人物や障害物により隠れてしまうことを防ぐため，上方に設置されることが多い．人物を上方から写した画像は，頭により手や足が隠されることで，特に自己隠蔽が発生しやすい上，見え方の変化の度合いも大きい．監視カメラの画像に対して人物姿勢推定を頑健に行うためには，これらの問題への対処が必要不可欠となる．

近年，Kinect[1] などの深度センサーを搭載したデバイスの普及により，深度画像を取得することが容易になってきている．深度センサーを搭載した監視カメラを設置すれば，街中で人物の写った深度画像を取得できるようになる．深度画像は，一般的に用いられる RGB 画像と比べ，人物の服装や照明の変化の影響を受けにくいだけでなく，画像に写っている物体の 3 次元位置を容易に求めることができ，被写体となる人物のプライバシーに関わる顔などの情報が取得されないという利点がある．これらの利点により，深度画像を用いた人物姿勢推定は，前述したプライバシーの問題や，服装，照明による多様性の問題に対処できるため，本研究では深度画像による人物姿勢推定を行う．また，視点の変化に対応するため，多様な姿勢をとらせた人物を様々な角度から撮影した深度画像を用意する．このような画像を用意すると，画像の総数が膨大になるので，大量のデータを効果的に学習できることで知られている畳み込みニューラルネットワーク (CNN) で学習を行う．様々な視点の深度画像からそれぞれ特徴を学習することで視点変化の問題に対処する．また，人物の関節位置の推定を深度画像から直接行うのではなく，人物を 31 部位に分け，一旦人物の深度画像の各画素が人物のどの部位に属するかを推定する部位推定を行い，その情報を用いて関節位置を推定する．この手順で関節位置推定を行うと，部位推定の時点で関節が写っているかどうかが判定できるので，関節に自己隠蔽が発生しているかどうかがわかる．

以下，本稿での構成を述べる．2 章で RGB 画像及び深度画像を用いた場合の

人物姿勢推定の関連研究を示し，それらの利点，欠点を述べる．3 章で提案手法の詳細を説明する．4 章で提案手法の有効性を示す実験を行い，さらに従来手法と関節位置推定精度を比較する．5 章で本研究の結論と今後の展望を述べる．

第 2 章　関連研究

人物姿勢推定では RGB 画像を用いた手法と深度画像を用いた手法が存在する．RGB 画像を用いた姿勢推定手法で代表的なものは Cao ら [2] によるものである．RGB カメラは広く普及しているので，実世界で撮影した写真や動画に容易に適用できるという利点がある．しかし，色の情報しかない RGB 画像のみから 3 次元位置を計算することは難しく，Cao らの手法により推定される人物の関節座標は画像中の 2 次元平面 (xy 平面) での位置である．3 次元座標を求めるためには Martinez ら [3] の手法のような 2 次元座標から 3 次元座標を求める手法を用いなければならないが，このような手法は 3 次元位置を計算して求めているのではなく，学習ベースの手法で 3 次元位置を推定しているだけである．手法を適用する際に正しい z 座標値が得られない可能性があるという点で，RGB 画像を用いた関節の 3 次元位置推定には限界がある．

深度画像を用いた人物姿勢推定で代表的なものは Shotton ら [4] による手法である．Shotton らは，3DCG ソフトなどを用いて十分な量の学習データを用意しており，姿勢推定に人物を 31 部位に分け，画像中の各画素が人物のどの部位または背景に属しているのかを推定し，その推定結果を元に関節位置の推定を行なうという手法をとっている．しかし，Shotton らの手法では学習データが人物を正面から写した画像のみであるため視点の変化に対応していない．

また，Haque ら [5] による手法では，CNN を用いて，人物を正面から写した画像を学習し，正面以外の視点の画像の関節推定をする際は，画像から得られる 3 次元点群を変換して正面に合わせることで視点の変化に対応している．各関節位置と同時に関節が見えているかどうかのマスクを学習するマルチタスク CNN の手法をとり，マスクにより見えていないと判断された関節の関節位置をダミー値とし，損失関数に影響しないように損失関数を設計することで自己隠蔽による悪影響を防いでいる．しかし，この点群変換は関節の局所領域ごとに行なわれており，関節同士の位置関係などの大域的な情報が失われている．このような大域的な情報も学習できるようにしておく方が関節推定精度は向上し

やすいと考えられる．

第3章　視点不変な人物姿勢推定

3.1　概要

本研究では，視点の変化に対応するため，人物を n 方向から仮想的に撮影した合成データ画像を生成し，それを学習データとして用いる．学習データ数が人物の姿勢の数×視点の数 (n 方向) と膨大な数となるので，大量データを効果的に学習できる畳み込みニューラルネットワーク (CNN) を学習に用いる．Shotton ら [4] の手法に倣い，人物を 31 部位に分け，人物の写った深度画像から各画素が人物の部位または背景の 32 クラスのうちのどれに属するかの尤度を推定するように学習する．人物を撮影した距離画像と，距離画像の各画素に対し 31 部位のどれに対応するかのラベル（部位ラベル）を付与した部位ラベル付き画像のペアを学習データとして CNN で学習する．このようにして学習した CNN により深度画像から人物の部位を推定し，その推定結果を用いて関節位置推定器により関節の 3 次元位置を推定する．なお，人物部位は図 1 のように分け，このような部位ラベル画像を元に図 2 のような関節位置を求める．このように，一旦部位ラベル推定を経由してから関節位置推定を行うことで，正解の部位ラベル画像からどの関節に自己隠蔽が発生しているかが容易に判断できることから，Haque ら [5] のように深度画像や正解の関節位置とは別に関節が見えているかどうかのマスクを CNN に入力する必要がない．さらに，関節位置を直接回帰するなどの手法を用いて推定した場合に比べて，関節の存在範囲を絞ってから関節位置を求められるため高速であるという利点もある．以降，各部位ラベルの名称として，左半身の部位には l，右半身の部位には r の文字をつけ，頭は頭部に u，顔面に w の文字をつけ，腕と足には付け根に近い部分を u，先端に近い部分を w の文字をつけたものを用いる．全ラベルの名称は，ruHead，luHead，rwHead，lwHead，neck，rChest，lChest，rWaist，lWaist，rShoulder，lShoulder，ruArm，luArm，rwArm，lwArm，rElbow，lElbow，rWrist，lWrist，rHand，lHand，ruLeg，rwLeg，luLeg，lwLeg，rKnee，lKnee，rAnkle，lAnkle，rFoot，lFoot，Background である．部位ラベルから正確な関節位置を求めやすくするため，Elbow や Knee など，ちょうど人物の関節にあたる部位のラベル面積は小さめに設定している．また，以降の関節画像では，画像によって関節を示す

線の太さを変えているが，これは関節がどこにあるのか見やすくするためであり，関節の推定精度の差などを表すものではない．

図 1: 人物部位ラベルの例

図 2: 人物関節位置の例

3.2　入出力

入力は背景差分済みの人物の写った深度画像とし，これを CNN に入力する．CNN は，入力画像中の各画素において，各部位の存在確率を算出し出力する．これにより得られた部位ラベル推定マップを関節位置推定器に入力することで各関節の推定 3 次元位置が得られる．入出力のフローチャートを図 3 に示す．

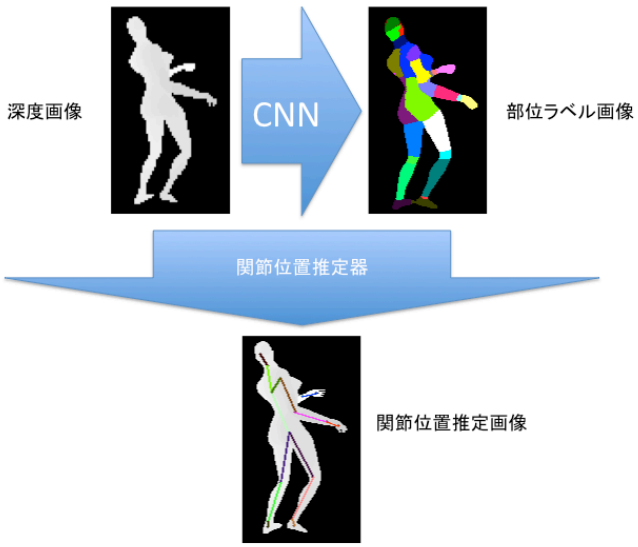

図 3: 提案手法の入出力フローチャート

3.3　学習データの生成

学習データの生成には高木ら [6] と同様の手法を用いる．3DCG モデル生成ソフト (Poser[7]) を用いて合成画像を作成し，これを学習データとすることで，十分な量の学習データ生成を可能にする．Poser で姿勢のコントロールが可能な人物の 3 次元形状モデルを生成し，日常生活で取りえない姿勢や，他のものと酷似した姿勢を取り除いたモーションキャプチャデータにより様々な姿勢を取らせる．この人物をランダムに回転させて床上に固定し，n 方向から撮影する．簡単のため，カメラの 3 次元回転はロール，ピッチ，ヨーのうちピッチ (上下方向) の回転のみ発生するものとし，ピッチはカメラに人物が写る範囲内でランダムに決定する．カメラ位置パターンは図 4 に示すような範囲で決定する．この人物を生成する際に人物の部位ごとに色分けしているので，この人物を撮影すれば，深度画像に加えて 31 部位のどれに対応するかのラベル（部位ラベル）を画素ごとに付与した部位ラベル付き画像が得られる．学習データは，このようにして得られた深度画像と部位ラベル付き画像のペアからなる．なお，人物は画像中に一人のみ写っており，人物領域以外の背景の部分には何も写っていないものとする．そのため，実画像を提案手法に適用する際はあらかじめ背景差分を行う必要がある．

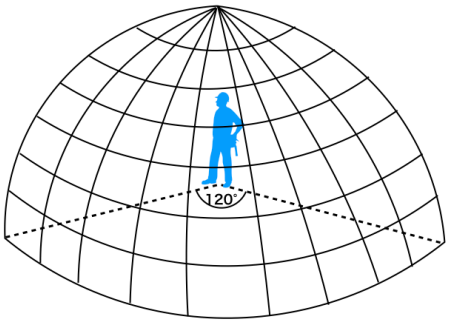

図 4: カメラ位置パターンモデルの例

3.4 CNN による部位ラベル推定の学習

3.4.1 CNN で学習する内容

以下では，学習データの画像は高さ h ピクセル，幅 w ピクセルの $h \times w$ サイズとし，人物の部位の数を 31 とする．学習データのうちの部位ラベル付き画像 (サイズ：$h \times w$) から，画像中の各画素においてその画素に各部位または背景が存在しているかどうかの 2 値の 32 次元 one-hot ベクトル (以下，これを $label$ と呼ぶ．サイズは $h \times w \times (31+1)$) が得られる．CNN は，学習データの背景差分済みの深度画像から，このベクトルを出力とするように学習することで，各画素において各部位または背景の尤度 (以下，これを $pred$ と呼ぶ．サイズは $h \times w \times (31+1)$) が得られる．

3.4.2 損失関数

categorical crossentropy とは，CNN の損失関数の一種であり，以下の式で表される．

$$E(pred) = -\sum \sum_{i,j=1}^{h,w} (label_{ij} \times \ln pred_{ij})$$

ここで，$label_{ij}$ は各画素にどの部位ラベルが付いているかの one-hot ベクトルであり，$pred$ は全画素における各部位ラベルの尤度であり，$pred_{ij}$ は各画素における各部位ラベルの尤度である．CNN は $E(pred)$ が最小となるように学習を行う．categorical crossentropy は，CNN で 3 つ以上のクラス分類を行う際に一般的に用いられている基本的な損失関数のひとつである．この式の特徴として，CNN の出力 ($pred$) のうち，各画素において正解のラベルに対して CNN が出力した尤度の数値のみが損失関数に影響し，各画素の正解でないラベルに対する CNN の出力した尤度の数値は影響しない，というものがある．すなわち，画像内の正解ラベルの数が多いラベルほど，そのラベルの学習が進みやすくなる．本研究においても，損失関数に categorical crossentropy を用いることができる．しかし，本研究の CNN の損失関数に categorical crossentropy を用いた場合，前述の特徴に起因する 3 つの問題が考えられる．

問題 1：人物領域と背景の面積比が大きい

人物領域の 10 倍から 20 倍もの面積を有する背景画素での損失が値の多くを占める．これにより，CNN は背景に重点をおいた学習を行ってしまうた

め，本来の目的である部位ラベルの学習の進度が遅くなり，ラベル推定精度が低下する．

問題 2：面積の小さいラベルに対する頑健性が確保されにくい

関節と対応する部位ラベルは面積の小さいものが多く，ラベルの欠損が発生してしまいやすい．このような欠損が発生すると，欠損したラベルに対応する関節位置の推定ができなくなる．

問題 3：自己隠蔽により見えてないラベルがほとんど学習されない

画像内に存在しないラベルが損失関数に与える影響は 0 であり，自己隠蔽が発生している画像において，見えていないラベルは他の見えているラベルの学習が進むことによる間接的な学習しかされない．

本研究では，categorical crossentropy をベースにし，$pred$ に重み付けを施すことで問題 1 及び問題 2 に対処する．重みの値は各部位ラベル及び背景ラベルごとに決定する．以下に重み $W_k(k = 1, 2, ..., 32)$ の計算式を示す．ここで，部位ラベル付き画像 (正解画像) における各部位ラベル及び背景ラベルの画素の面積を $area_k(k = 1, 2, ..., 32)$ とする．$k = 32$ は背景ラベルであり，$h \times w$ は画像全体の面積である．

部位ラベル

$$W_k \propto 1 - \frac{area_k}{h \times w} \ \ and \ \ \frac{\sum_{k=1}^{31} area_k \times W_k}{area_{32}} = 1 \ (k = 1, 2, ..., 31)$$

背景ラベル

$$W_{32} = \frac{\sum_{k=1}^{31} area_k}{h \times w}$$

この重み式の特徴は，背景領域と人物領域の面積の逆比をそれぞれ背景ラベル及び各部位ラベルをかけることで，背景ラベルの損失が損失関数に与える影響を小さくしており，さらに，人物領域内の各部位ラベル間においても面積の小さいラベルほど重みを大きくして損失関数に与える影響を大きくしている．面積が 0 の部位ラベルが最大の重みとなる．このようにして計算された重みを用いて，以下のような損失関数を考える．

$$F(pred) = -\sum_{i,j=1}^{k}\sum^{h,w} W_k(label_{ijk} \times \ln pred_{ijk} + invisible_k \times \ln(1 - pred_{ijk}))$$

式中の $invisible_k$ は，部位ラベル $k(k = 1, 2, ..., 32)$ が画像中に存在するなら 0，存在しないなら 1 となる変数である．見えていないラベル k に対する $pred_{ijk}$ の値が全て 0 になるよう学習させることで問題 3 に対処する．しかし，この項を付け足すことで，逆に自己隠蔽が発生していない部位を自己隠蔽していると誤推定しそのラベルが欠損してしまうという可能性がある．自己隠蔽に対処できることよりも，欠損が発生しにくくなるようにすることを重視する場合，この項を除いた以下の損失関数が考えられる．

$$G(pred) = -\sum_{i,j=1}^{k}\sum^{h,w} W_k \times label_{ijk} \times \ln pred_{ijk}$$

なお，$F(pred)$ 及び $G(pred)$ は，面積の小さい部位ラベルの学習が進みやすい一方，面積の大きい部位ラベルの学習が進みづらく，面積の大きい部位ラベルでは推定精度の低下が考えられる．しかし，両手及び両足の関節位置推定は全て小さな部位ラベルを用いて行なっているため，部位ラベル推定精度低下が与える手足の関節位置推定精度への直接的な影響はない．さらに，頭や胴体の部位ラベルは推定が容易であるため，多少識別性能が下がっても部位ラベル推定精度にそれほど影響はない考えられる．本研究では，以下の実験において，$E(pred)$，$F(pred)$，$G(pred)$ をそれぞれ用いた CNN のラベル推定精度や自己隠蔽への対応能力を調べる．

3.4.3 CNN のネットワーク構成

セマンティックセグメンテーションとは，画像中の各画素において，その画素がどのオブジェクトに属するかのラベル付けを行うというものである．CNN でセマンティックセグメンテーションを行う際には，畳み込み層とプーリング層のみで構成されており，end-to-end で高速な学習ができる fully convolutional network(FCN) を利用するのが適しているとされている [8]．FCN には全結合層がないため，遠い画素間での情報は共有されない．一般的なセマンティックセグメンテーションでは，ある領域のラベルを決定する際にそこから遠く離れた

領域の情報は必要ないため，FCN をセマンティックセグメンテーションに問題なく適用できる．本研究の部位ラベル推定手法もセマンティックセグメンテーションの一種であるが，人物の部位は画像中に一箇所ずつにしか存在し得ないため，同じ部位ラベルが複数領域に現れることはないという制約が存在する点が一般的なセマンティックセグメンテーションとの違いである．さらに，人物画像は右腕と左腕，右足と左足のように形状が酷似しており間違えやすい部位もあるため，局所的な情報のみでラベル推定を行う FCN を利用するとラベル推定がうまくいかない可能性がある．そこで，本研究では，FCN に全結合層を加え，局所的な情報だけでなく大域的な情報も学習できるようにする手法を提案する．これにより，CNN 内の総パラメータ数は大幅に増えるため，学習時間が増加するとともに，余分なパラメータが部位ラベル推定に悪影響を与える可能性がある一方，前述した制約を学習できるようになり，左右の手足を取り違えるケースが減ると考えられる．本研究では，以下の実験において，FCN をそのまま用いた場合と FCN に全結合層を追加した場合との部位ラベル推定結果を比較し，全結合層を追加することによる利点と欠点がどれほど部位ラベル推定に影響を与えるか考察する．なお，本研究で用いた CNN(全結合入り) のネットワーク構成図は付録の図 A.1 に記している．

3.5 関節位置推定器

関節位置推定器は，CNN に入力した深度画像と CNN が出力した $pred$ の 2 つを用いて各関節の 3 次元位置を推定する．入力された深度画像から，人物領域内の各画素においてその画素に写っている人物表面の 3 次元位置をあらかじめ求めておく．各関節の推定関節位置 $\hat{x}_j$ は，$pred$ からその関節に対応する部位ラベルの尤度を得て，その尤度を重み付けた 3 次元点群の重心のうち，もっとも確率の高いものとなる．3.1 で述べたように，関節に対応する部位ラベルは面積が小さいため，部位ラベル推定が正しくできたならば部位ラベルの重心が真の関節位置に十分に近くなる．この重心の探索は，Shotton ら [4] と同様に重み付き mean shift を用い，重心を求める前に，求める関節に対する $pred$ の値が 0.14 以下の場合には重みを 0 として扱う．すなわち，全ての画素で $pred$ の値が 0.14 以下である部位に対応する関節は見えていないものとし，推定を行わない．なお，このようにして求められた $\hat{x}_j$ は厳密な関節位置ではなく，カメラに写っている関節の人物表面上の 3 次元位置である．したがって，実際の関節位置 x_j を求め

るために，$\hat{x_j}$ をカメラの奥行き方向に移動させる必要がある．本研究では，高木 [6] の手法に倣い関節ごとの奥行き方向のオフセットを表 1 のように定めた．実際の推定関節位置 x_j は，$\hat{x_j}$ の奥行き方向に以下のオフセットの値を加えたものとなる．なお，表 1 には，関節に対応する部位ラベルも示している．

表 1: 関節ごとのオフセット及び対応する部位ラベル

関節	オフセット (cm)	部位ラベル
Head	10.0	ruHead luHead rwHead lwHead
Neck	7.0	neck
Chest	9.5	rChest lChest
Waist	9.5	rWaist lWaist
Shoulder	6.5	r(l)Shoulder
Elbow	4.1	r(l)Elbow
Wrist	2.6	r(l)Wrist
Hand	1.0	r(l)Hand
Knee	4.1	r(l)Knee
Ankle	4.1	r(l)Ankle
Foot	2.0	r(l)Foot

第 4 章　実験・評価

4.1　評価尺度

以下の実験の評価尺度を以下のように定める．

1. 以下の式 $P_k(pred)(k = 1, 2, ..., 31)$ を画像 1 つに対する各部位ラベルの推定精度とし，全テスト画像におけるこの値の平均を部位ラベル推定の評価尺度とする．式中の $\sum_{i,j=1}^{h,w} label_{ijk} \times pred_{ijk}$ が部位ラベル k が推定された領域の面積であり $area_k$ が部位ラベル付き画像 (正解画像) における部位ラベル k の面積である．

$$P_k(pred) = \frac{\sum_{i,j=1}^{h,w} label_{ijk} \times pred_{ijk}}{area_k}$$

2. 関節位置推定の評価尺度は，各関節において，CNN の出力である $pred$ から関節位置推定器により推定した関節位置と，正解の部位ラベル画像である $label$ から関節位置推定器により推定した関節位置 (すなわち，部位ラベル推定に基づく関節位置推定の理論上の限界性能) との (3 次元距離における) 誤差が一定値以下 (ここでは，Shotton ら [4] に倣い 10cm とする) となった関節および自己隠蔽を正しく認識できた関節の割合とする．

4.2 概要

本章では，提案手法の有効性を示すため，Long ら [8] による一般的なセマンティックセグメンテーションを行う手法を用いた場合の部位ラベル推定精度の比較や，Shotton ら [4] による従来の人物姿勢推定手法との関節位置推定精度の比較を行う．また，前章で述べた全結合層の有無や損失関数の違いでどれほど部位ラベル推定精度に差があるか調べ，もっとも精度の高くなる組み合わせを模索する．なお，以下の実験で用いた画像のサイズは学習データ，テストデータに関わらず全て縦 416 ピクセル，横 512 ピクセルとした．

4.3 損失関数の性能評価

提案手法の損失関数が部位ラベル推定に有効であることを示すため，提案手法の損失関数を用いた CNN により学習した場合と，セマンティックセグメンテーションのベースラインの手法である Long ら [8] の手法により学習した場合とで部位ラベル推定精度を比較した．ここで，単純な部位ラベル推定精度だけでなく，自己隠蔽が発生していても正しくラベル推定が行われているかも調べるため，学習データ，テストデータとして用いた深度画像は，全て自己隠蔽の発生しやすい上方視点からの画像 15100 枚 (うち学習データ 15000 枚，テストデータ 100 枚) とした．その結果を図 5 に示す．図中の青の棒グラフが従来手法 (損失関数：$E(pred)$) の部位ラベル推定精度，緑の棒グラフおよび赤の棒グラフが提案手法 ($F(pred)$(自己隠蔽対応を重視)，$G(pred)$(欠損対応を重視)) の部位ラベル推定精度である．ほぼ全ての部位ラベル推定精度で $E(pred)$ よりも $F(pred)$ および $G(pred)$ が上回っている．また，$F(pred)$ と $G(pred)$ の比較では平均では $G(pred)$ が上回っているが，$G(pred)$ が高いのは関節位置推定で使用しない腕 (Arm) や足 (Leg) が中心であり，下半身を中心に $F(pred)$ が大きく上

回っている部位もあるため，一概にはどちらが優れているかの判断がつかない．また，図 6 は，画像 100 枚中，各関節で自己隠蔽が発生している画像の数 (黄の棒グラフ) に対して，そのうち自己隠蔽が発生していることを正しく推定できた画像の数 (青 ($E(pred)$)，緑 ($F(pred)$)，赤 ($G(pred)$) の棒グラフ) を示している．自己隠蔽対応を重視した $F(pred)$ が多くの部位ラベルにおいてより高い対応力を発揮しているが，他の 2 つの損失関数とそれほど大きな差は見られなかった．逆に言えば，どの損失関数を用いても，平均で 60 パーセントから 70 パーセントほどは自己隠蔽に対応できている．部位ラベル推定を経由すれば，損失関数に特別な手を加えなくとも，ある程度自己隠蔽を考慮した関節位置推定が行えることがわかった．

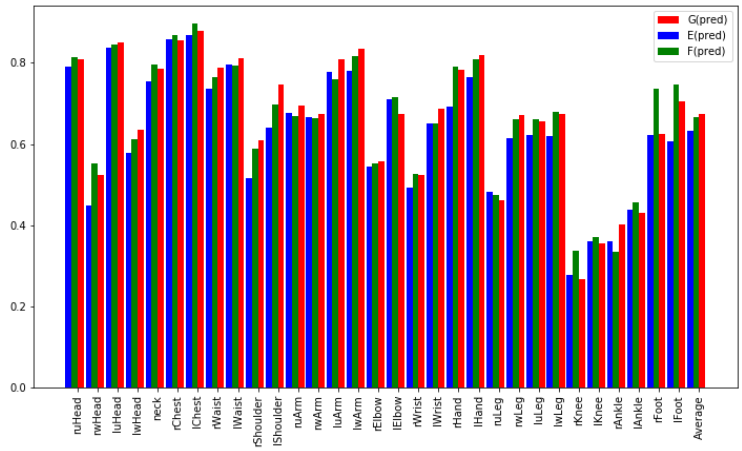

図 5: 損失関数別各部位の推定精度比較

4.4　全結合層の有無が部位ラベル推定精度に与える影響

3.4.3 で述べたように，部位ラベル推定は全結合層を入れることで，全体の精度が下がるものの，左右を間違えて複数箇所に誤推定してしまいやすい手や足の推定精度が向上すると考えられる．そこで，全結合層の有無以外の条件を全く同じにして学習した 2 つの CNN により部位ラベル推定精度を比較した．その結果を図 10 に示す．この図から，全結合層を入れても大きく推定精度が下が

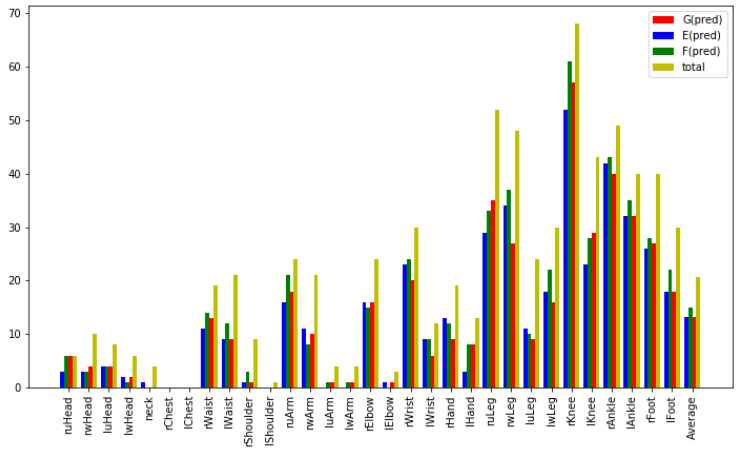

図 6: 損失関数別各部位の自己隠蔽対応力比較

ることはなかった．一方，図 7 のように全結合層なしだと左手の一部を右手と誤推定してしまう (どちらの手も黄色の lHand ラベルをつけている) が図 8 のように全結合層ありでは正しく推定できる，というケースがみられ，全結合層を入れることでこういった間違いは減ると考えられる．しかし，CNN の識別性能が高く，全結合層なしでも同じ部位ラベルを複数箇所に推定してしまうというケースはあまり多くないため，大きな推定精度向上には繋がらなかった．

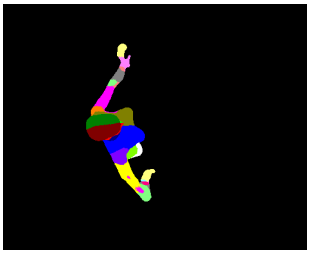

図 7: 全結合層なしの部位ラベル推定画像

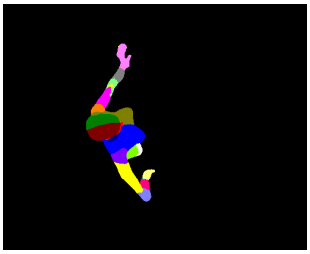

図 8: 全結合層ありの部位ラベル指定画像

図 9: 図 7 と図 8 の正解の部位ラベル画像

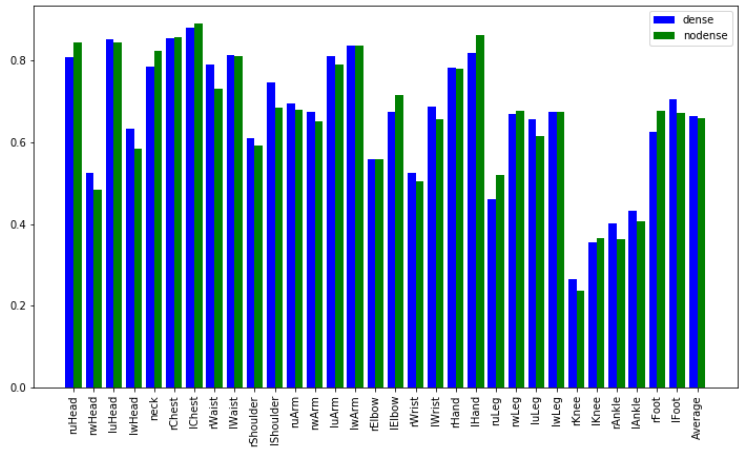

図 10: 全結合層の有無による各部位の推定精度の比較

4.5 視点を変化させた場合の各損失関数の性能比較

これまでの実験では，簡単のために視点を固定した状態で比較してきた．本実験から，視点を固定せず 63 方向の視点を学習データに含める．本実験では，従来の損失関数 $E(pred)$ と，提案した損失関数 $F(pred)$ と $G(pred)$ をそれぞれ用いた場合の部位ラベル推定精度を比較した．その結果を図 11 に示す．この結果から，自己隠蔽に対処しようとした $F(pred)$ がもっとも精度が低くなっているが，これは自己隠蔽が上方視点などの特殊な視点でしか発生しづらく，自己隠蔽のない画像の部位ラベル推定では自己隠蔽を評価する項が悪影響を与えているからだと考えられる．自己隠蔽への対応力の差を考慮しても，視点を固定しない場合は $G(pred)$ の方が良い性能を示すことが明らかとなった．

4.6 視点の変化に対する頑健性

前述のように，カメラの位置によって人物の見え方は大きく変化する．全ての視点の深度画像を学習データとした場合，画像によって人物の部位の形状が大きく違うため，別視点どうしの画像の情報が衝突し，うまく学習が行われない可能性がある．ここで，ある視点から撮影した深度画像のみを学習データとしてその視点に特化した CNN と，様々な視点から撮影した深度画像を学習データとしている提案手法が同一のテスト画像に対して同等の推定精度となれば，

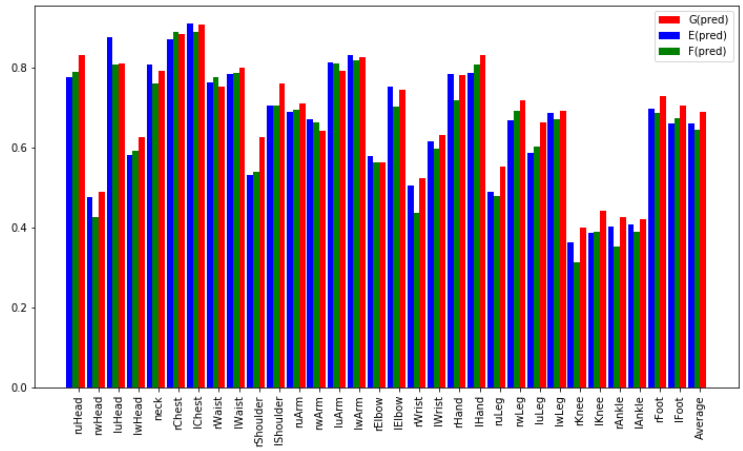

図 11: 各部位の推定精度の比較

提案手法は様々な視点の画像を正しく学習できていると言える．そこで，学習データ以外の条件を統一 (全結合層あり，損失関数 $G(pred)$) し，学習データを上方のある 1 点のみとした場合 (総数 15000) と 63 方向の視点とした場合 (総数 15000 × 63) とで部位ラベル推定精度及び関節位置推定精度を比較した．その結果を図 12 及び図 13 に示す．図中の青の棒グラフ (only) が学習データの視点を固定した場合で，緑の棒グラフ (all) が学習データの視点がランダムな場合である．この結果から，全ての視点を学習データとしても推定精度に大きな低下はなく，部位ラベル推定精度では膝を中心に精度が向上し，それに伴って関節位置推定精度は大幅に向上するなどの良い影響が見られた．これは，上方視点からだと膝部分の情報が得にくいため，視点の変化による人物の見え方の変化がそれほどなく，膝部分が見えやすい近くの視点の情報を推定に利用したからだと考えられる．一方，自己隠蔽に対する頑健性では図 14 のように視点固定の場合がやや高いが，それほど大きな差は見られなかった．次に，全視点を学習した CNN を実際に適用した例を示す．図 15 から図 20 は同じ姿勢をとった人物を正面から撮影した場合（上段）と上方から撮影した場合（下段）の深度画像に提案手法を適用した結果であり，左から入力の深度画像，部位ラベル推定画像，関節位置推定画像である．この結果から，カメラの位置を変化させてもラベル推定及び関節位置推定が可能であるということがわかる．

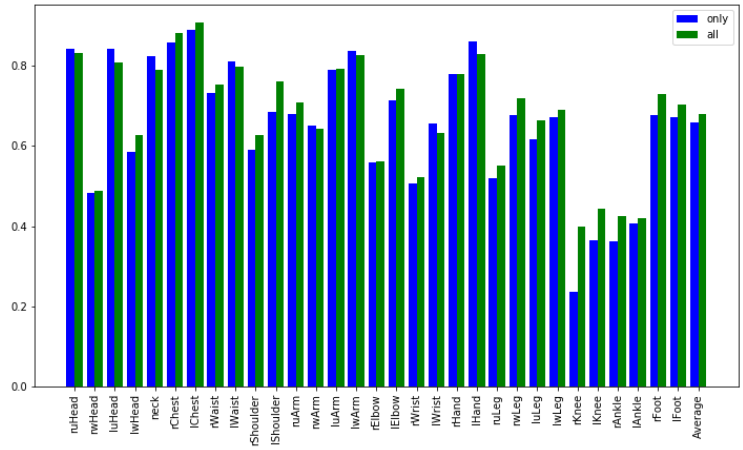

図 12: 学習データ別各部位の推定精度比較

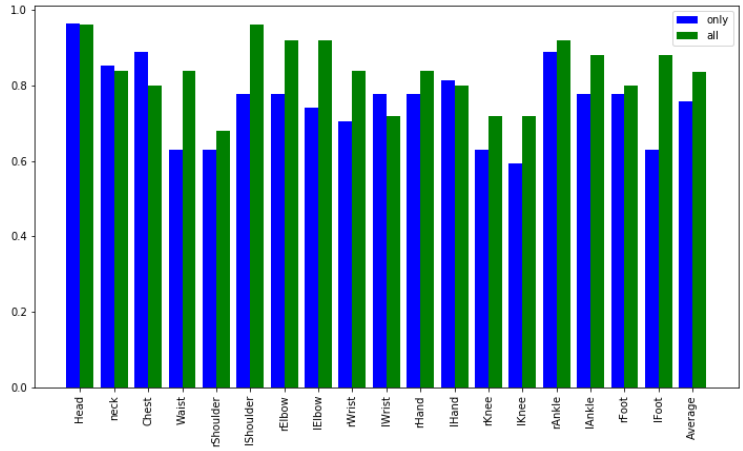

図 13: 学習データ別各関節位置推定精度比較

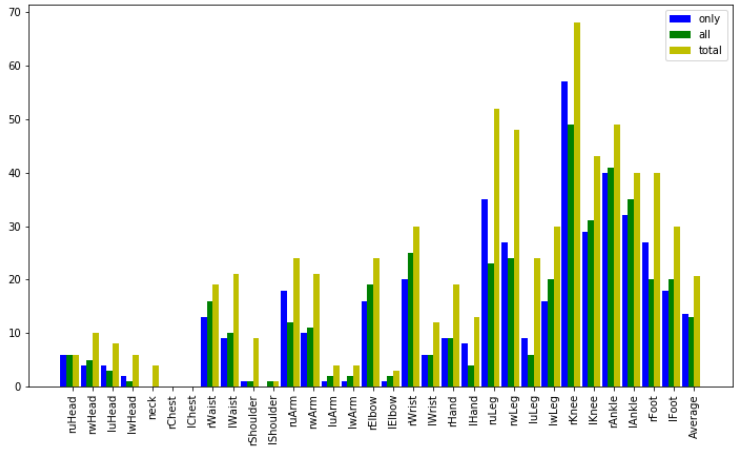

図 14: 学習データ別自己隠蔽対応力比較

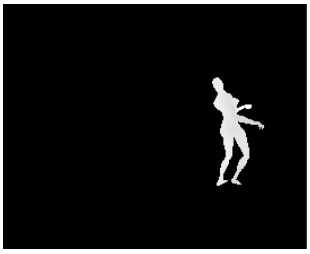

図 15: 正面から撮影した場合の深度画像

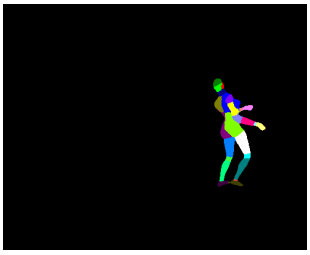

図 16: 正面からの部位ラベル推定画像

図 17: 正面からの関節位置推定画像

図 18: 上方から撮影した場合の深度画像

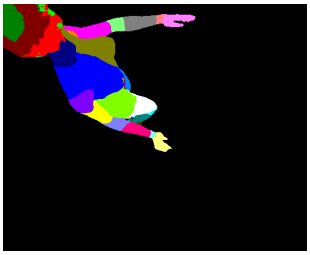

図 19: 上方からの部位ラベル推定画像

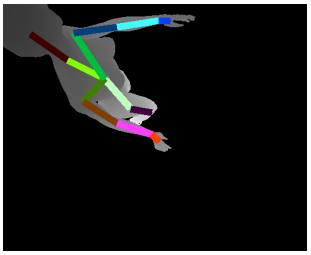

図 20: 上方からの関節位置推定画像

4.7　関節位置推定における従来手法との性能比較

本実験では，提案手法と，Shotton らによる従来手法との関節位置推定精度を比較する．なお，Shotton らの手法は詳細なプログラム，データセットが公開されていないため，論文中で示されたデータから本実験の評価尺度に合わせた関節位置推定精度を計算した (なお，Shotton らは胸および腰の位置推定は行なっていない)．提案手法では学習データは 63 視点の画像を 15000 × 63 枚を用いテストデータも視点ランダムにした画像 100 枚とした．その結果を図 21 に示す．ほぼ全ての関節で大幅な推定精度上昇が見られた．なお，Shotton らの数値は正面からの画像のみをテストデータとした場合のものであり，提案手法と同様に様々な視点の画像をテストデータとした場合はこれらの数値よりさらに低くなると考えられ，その点を考慮しても提案手法がより高い精度となっている．

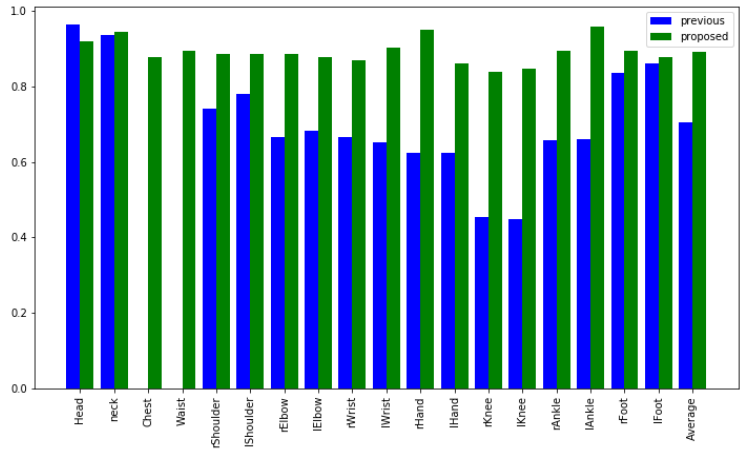

図 21: 関節位置推定精度の比較

4.8　実データへの適用可能性

これまでの提案手法の性能評価は全て合成データを用いて行ってきたが，実用化には実データに対する頑健性が重要になる．実画像から部位ラベル画像や正解の関節位置画像を生成することが難しいため，本実験では提案手法を実画像に適用し，目視による確認にとどめる．深度センサーつきカメラにより人物

を撮影した実画像 (図 22) に背景差分を施す (図 23)．なお，背景差分には背景画像との差分をとった画像を固定閾値処理し，オープニングを施したマスクにより前景を抽出する手法を用いた．この画像を CNN に入力し，出力として得られた部位ラベル画像が図 25 である．図 23 で示されるように，背景差分時に発生したノイズを人物領域の一部であると誤推定している．本研究で用意した学習データは，全て背景を取り除いた画像であり，本来人物領域外の領域で画素値が与えられることを考慮しておらず，そのような領域に対して何かしらの部位ラベルを与えてしまう．そのため，このような推定結果になると考えられる．また，足は背景 (床) との距離が近いため，背景差分時にどうしても欠損が発生してしまい，背景差分を前提とした場合足のラベル推定を行うのは難しい．この推定結果を関節位置推定器に入力した場合の出力が図 26 であるが，ノイズの影響により正しい推定ができていない．また，ノイズを発生させることなく背景差分ができた場合を仮定し，背景差分時に発生するノイズを全て取り除いた状態で部位ラベル推定，関節位置推定を行なった画像を付録の図 A.2，図 A.3 に示しておく．仮にノイズの発生しない背景差分を行うことができても，合成画像にはない実画像特有の服のわずかなシワの変化やカメラのキャプチャ精度の低さなどが推定に悪影響を与えていると考えられる．実画像に対して提案手法を適用できるようにするためには，このような実画像特有の変化を再現した合成画像を生成する，実画像を大量に集めるなどして実画像認識に特化した学習データを用意する必要がある．

図 22: 深度カメラによる実画像

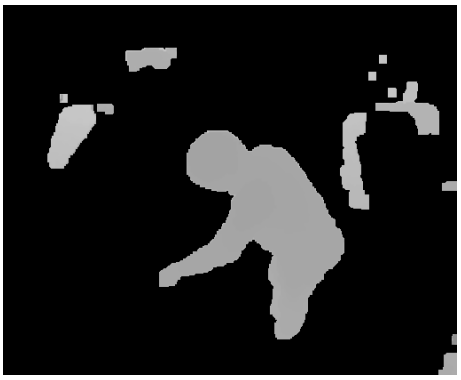

図 23: 背景差分画像

図 24: 部位ラベル例

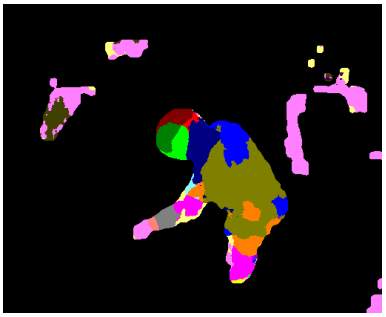

図 25: 部位ラベル推定画像

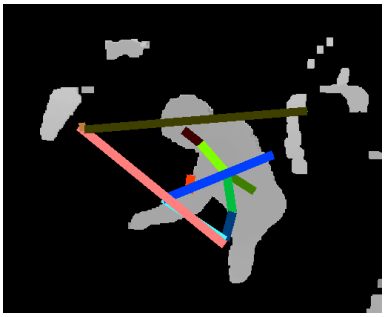

図 26: 関節位置推定画像

図 27: 関節位置例

4.9 各実験に対する考察

本研究では，$F(pred)$ と $G(pred)$ の 2 つの損失関数を提案したが，4.3 と 4.5 の実験から，$F(pred)$ は自己隠蔽の対応能力が高い一方，自己隠蔽が少ない学習データでは精度が下がりやすく，$G(pred)$ を損失関数として採用するのがよいと考えられる．また，4.6 の実験において，学習データを視点ランダムにしても推定精度が下がらなかったのは，同じ姿勢を様々な視点で撮影した画像全てを学習データとするのはデータオーギュメンテーションの一種と捉えることができ，これにより未知の画像に対応しやすくなったためだと考えられる．すなわち，ある視点に対する別視点の画像は，推定に悪影響を与えるのではなく，むしろ推定の補助になっていると言える．なお，データオーギュメンテーションでの精度向上が発生しないように，視点ランダムな学習データの総数を 15000 × 63 枚ではなく各姿勢からランダムに 1 視点のみとり出した画像 15000 枚を学習データとして学習した場合，1 視点あたりの画像数が約 240 枚と少ないため，CNN の性能を十分に発揮できず，部位ラベル推定精度はあまり高くならない，

また，視点の変化に対応していることをより明確にするため，CNN に部位ラベルの推定と同時に画像がどの視点から撮影されたものかを学習させるという方法が考えられる．これにより，視点によって個別に特徴を学習でき，部位ラベル推定精度の上昇が見込めるだけでなく，視点が上方だと推定された場合は $F(pred)$ を用いるというような損失関数のさらなる改良が可能となる上，後述する関節位置推定器でのオフセット値決定にも役立てることができる．

4.10　関節位置推定器に関して

前述の通り，評価尺度として用いた関節位置は，正解の部位ラベル画像から関節位置推定器によって求めたものであり，真の関節位置ではない．正解の部位ラベル画像から関節位置推定器により求めた関節位置と真の関節位置との 3 次元位置での誤差が 10cm 以内になる関節の割合は 91.4 パーセントである [4]．すなわち，CNN がどれほど頑健に部位ラベル推定を行なったとしても，関節位置推定機を改良しない限り，この値を超えることはできず，この手法の限界となっている．本実験の評価尺度では，関節位置推定器の性能評価はしていないものの，関節位置推定器の性能も良いとは言い難く，特に以下の 2 パターンで誤った関節位置推定をしてしまいやすい．

問題 1：首推定

上方からの視点の画像では，首は頭により隠されてしまう．図 30 のような，首（青色のラベル）が一部しか見えていない画像では，首の関節位置を頭の真下に推定することができず，図 31 のように首の関節位置が真の関節位置より大きくずれてしまう．

問題 2：胸腰の位置推定

胸や腰の位置は，それぞれ胸の部位ラベル，腰の部位ラベルから求めている．図 28 のような，左腰（紫色のラベル）の一部が手によって見えない画像では，腰ラベル群の極大点が右に偏ってしまい，図 29 のように腰の関節位置が真の関節位置 (rWaist ラベルと lWaist ラベルの境界の中間あたり) より大きくずれてしまう．

Shotton ら [4] の手法では，視点が正面固定なので問題 1 は発生せず，問題 2 は，部位ラベル推定の精度が低くても関節位置を大きく外さないようにするためにこのような関節位置推定をしていると考えられる．視点ランダムな画像からより正確な関節位置を求めるには，Shotton らの関節位置推定器をそのまま用いるのではなく，これらの問題に対処できるように改良を加える必要がある．例えば，カメラの位置によってオフセットの値を変えることで問題 1 に，胸や腰の位置推定では，提案手法の部位ラベル識別性能の高さを生かし，lChest と rChest，lWaist と rWaist のエッジ部分の中央あたりを推定関節位置とすることで問題 2 に対処できると考えられる．(Shotton らのラベル推定性能では推定した部位ラベルのエッジが荒く，このような関節位置推定手法をとるのは難しい．)

図 28: 左腰の一部が見えない部位ラベル画像 (拡大処理済)

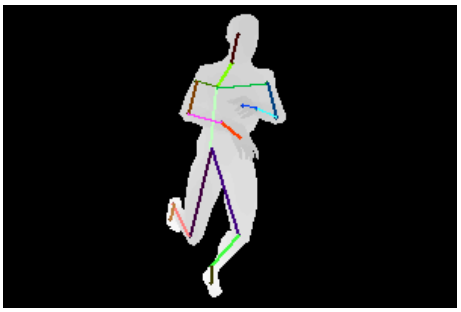

図 29: 図 28 を入力とした関節位置推定器の出力画像 (拡大処理済)

図 30: 首の大部分が頭に隠されている部位ラベル画像 (拡大処理済)

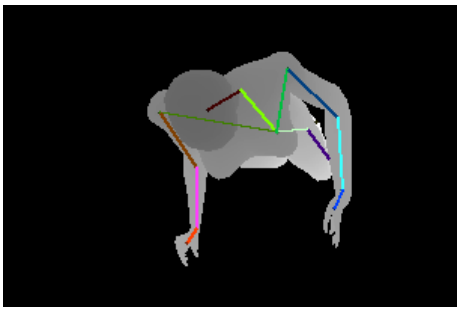

図 31: 図 30 を入力とした関節位置推定器の出力画像 (拡大処理済)

4.11 自己隠蔽の発生に関して

本研究では，自己隠蔽を考慮せずに推定した時に，全く違う場所を推定してしまうことによる損失を重視し，自己隠蔽を考慮した推定を行えるような手法を目指した．右腕全体が隠れているというようなケースでは，このような手法をとるのがよいと考えられるが，撮影角度の関係で関節にあたる部位のみが見えていないようなケースでは，無理に関節位置推定をしても，もし隠れていなければその位置に推定されたと思われる位置に推定できることもある．その例が図 32 である．本来なら図 33 に示すように部位ラベル画像上は右足の膝は隠れているが，右足全体はほとんど隠れていない．図 32 は本研究の評価尺度では誤推定になるが，右膝の推定関節位置は真の関節位置にかなり近いと思われる．すなわち，自己隠蔽の発生を完全に判断できるようになるのは，必ずしもメリットになるとは限らないのである．このように，自己隠蔽の発生を全てひととおりに考えるのではなく，どういうパターンの自己隠蔽なのかを分けて考え，無理に推定しても問題ないケースなら推定を行う，というように臨機応変な判断

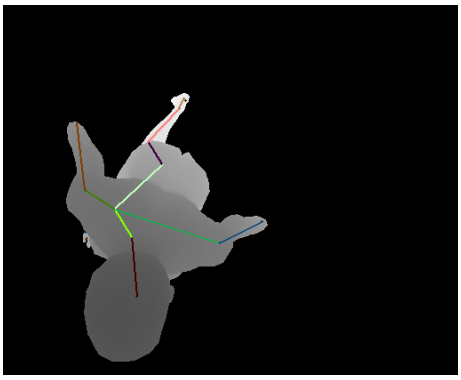

図 32: 右膝が隠れていないと誤推定した画像

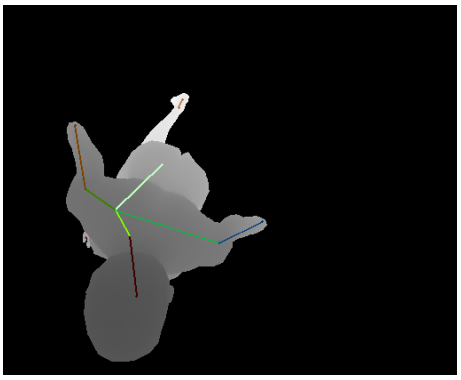

図 33: 図 32 の正解の関節位置画像 (本来なら右膝は隠れている)

ができるようにすると，小さな隠れに対して強くなると考えられる．

また本研究では，自己隠蔽の発生している関節は推定しない手法としている．しかし，実世界に適用する場合は見えていない関節の 3 次元位置も推定できるような手法であることが望ましい．特定の条件下ならば，部位ラベル推定画像からでも，自己隠蔽されている関節位置を推定できる．例えば，上腕と前腕が見えていて，肘のみが頭に隠されて見えていない場合，肘は上腕と前腕のちょうど中間あたりだと推定できる．このような手法をとれば，現在関節位置推定に使用していない腕や腿のラベルも関節位置推定に使用できるようになる．また，見えていない関節の位置を他の見えている関節や部位の情報を利用して求める以外にも，Sun ら [9] の手法のような，推定した関節位置を修正する手法を適用すれば，自己隠蔽の発生により誤推定してしまった場合も正しい位置に修正できるので，自己隠蔽に対応した関節位置推定ができると考えられる．

第 5 章　結論

本研究では，人物の写っている深度画像から，画像内に写っている人物の関節の 3 次元位置を推定する手法を提案し，従来の部位ラベル推定に基づく人物姿勢推定手法よりも高い関節位置推定精度を達成し，視点の変化に対して頑健に推定することができた．さらに，部位ラベル付き画像では，関節に自己隠蔽が発生していた場合，同時にその関節に対応した部位ラベルも隠れる．そのため，部位ラベル推定を経由することで同時に関節の自己隠蔽の有無を推定でき，自己隠蔽をある程度考慮した手法であることが示せた．

また，人物姿勢推定を実世界で適用するには，実画像に対しても頑健に関節位置を推定できなければならないが，提案手法で用いた学習データは実画像で発生するノイズを考慮していない．そのため，ノイズに対する頑健性が低く，現状では実画像に提案手法を適用することは難しい．実画像に対して頑健な人物姿勢推定を行える手法に改良することが今後の課題である．

謝辞

本研究を進めるにあたり，熱心な御指導を賜りました美濃導彦教授，飯山将晃准教授に深く感謝致します．また，本研究全体を通して親身になって御指導を賜りました橋本敦史助教，薗頭元春氏に心より感謝致します．最後に，研究生活を楽しいものにして下さった美濃研究室の皆様に感謝致します．

参考文献

[1] Microsoft: Kinect.

[2] Z.Cao, T.Simon, S. Y.: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields, *CVPR* (2017).

[3] J.Martinez, R.Hossain, J.Romero and J.Little: A simple yet effective baseline for 3d human pose estimation, *CoRR* (2017).

[4] R.Girshick, J.Shotton, P.Kohli, A.Criminisi and A.Fitzgibbon: Efficient regression of general-activity human poses from depth images, *IEEE,ICCV* (2011).

[5] A.Haque, B.Peng, Z.Luo, A.Alahi, S.Yeung and L.Fei-Fei: Towards Viewpoint Invariant 3D Human Pose Estimation, *CVPR* (2015).

[6] 高木和久: カメラ位置毎の学習に基づく視点不変な人物姿勢推定, 京都大学大学院情報学研究科修士論文 (2017).

[7] SmithMicro: Poser Pro 2014.

[8] J.Long, E.Shelhamer and T.Darrell: Fully Convolutional Networks for Semantic Segmentation, *CVPR* (2015).

[9] K.Sun, C.Lan, J.Xing, W.Zeng, D.Liu and J.Wang: Human Pose Estimation using Global and Local Normalization, *ICCV* (2017).

付録：CNN のネットワーク構成

以下に，本研究で用いた CNN のネットワーク構成を示す．

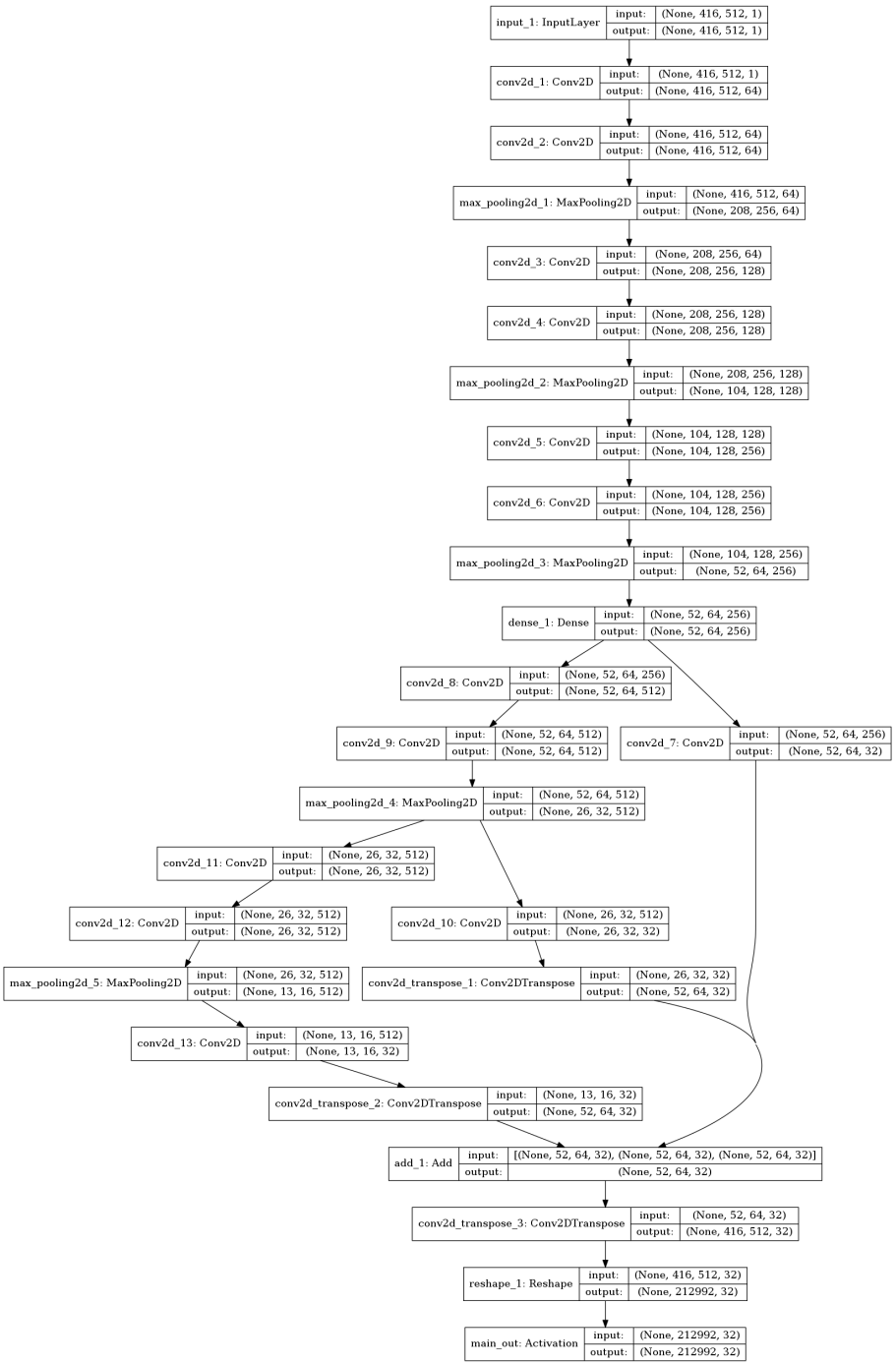

図 A.1: ネットワーク構成 (画像サイズ 416 × 512 の場合)

以下に，4.10 で述べた，実画像にノイズのない背景差分を施した場合の部位ラベル画像，関節位置推定画像を示す．

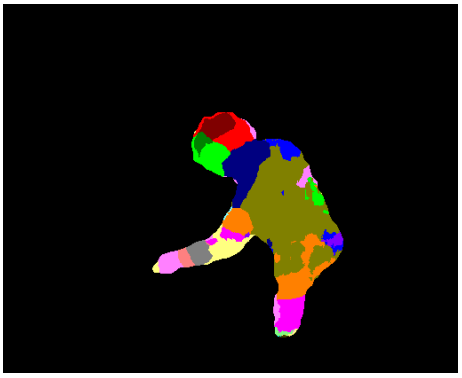

図 A.2: ノイズのない背景差分を施した場合の部位ラベル推定画像

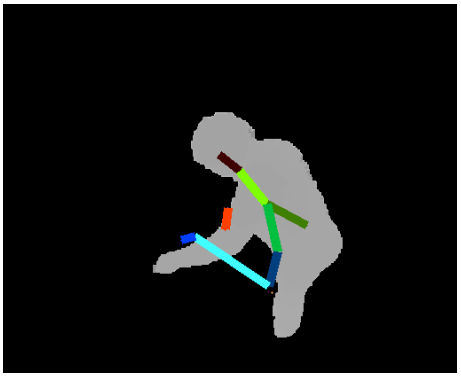

図 A.3: ノイズのない背景差分を施した場合の関節位置推定画像