5R-06

合成画像への画像変換による実画像人物姿勢推定

髙橋 龍平† 飯山 将晃‡ 薗頭 元春‡ 橋本 敦史††
京都大学大学院情報学研究科† 京都大学学術情報メディアセンター‡
オムロンサイニックエックス株式会社††

1. はじめに

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て人物姿勢推定を行えば，不審な行動をとる人物の自動検知や，買い物客が手に取るなどして興味を示した商品を把握することによるマーケティングに活用できる．人物姿勢推定とは，人物が写った画像からその人物の各関節の 3 次元空間中での位置を推定する問題のことである．

教師付き学習により姿勢推定を行う場合，教師データとして人物が写った画像とその人物の関節位置（アノテーション）のペアを用意する必要がある．実画像を教師データとして用いる場合，アノテーションを得るには特殊な環境が必要であり，高コストである．一方，3DCG モデル生成ソフトウェアを用いて生成した合成画像には容易にアノテーションが可能である．アノテーション付き合成画像を教師データとして畳み込みニューラルネットワーク（CNN）で学習した姿勢推定器は，合成画像に対しては十分な性能を発揮する(図 1(a))ものの，同じモデルをそのまま実画像に適用しても，合成画像には存在しないノイズや欠損などの影響により，うまく推定できない（図 1(b)）．そこで，本研究では，実画像を，ノイズや欠損などを取り除いた合成画像と類似する画像へ変換する手法を提案する．これにより，合成画像を用いて手動アノテーションなしで学習した姿勢推定器を用いて実画像に対する姿勢推定を行うことが可能となる．（図 1(c)）

(a) 合成画像　(b) 実画像 提案手法非適用　(c) 実画像 提案手法適用

図 1：各画像に対する姿勢推定器の推定結果

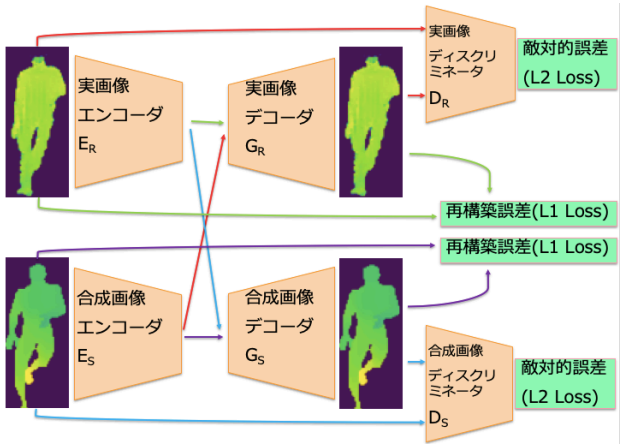

図 2：提案手法におけるネットワーク構成

2. 提案手法

本研究における画像変換では，変換前の画像（実画像）と変換後の画像（合成画像）のペアを用意できないため，CycleGAN[2]などに代表されるペアなし画像変換を用いることとする．提案手法においても，GAN を利用した画像変換ネットワークを用いる．提案手法で用いたネットワーク構成を図 2 に示す．画像が入力されたとき，その画像に対応するエンコーダ（実画像ならE_R，合成画像ならE_S）により特徴抽出を行い，デコーダにより変換画像の生成および入力画像の復元を行う．損失関数L_{total}は入力画像，変換画像，復元画像から計算される敵対的誤差，再構築誤差，変換誤差の線形和とする．以下にこれらの誤差の詳細を示す．エンコーダとデコーダを組み合わせたものを生成器と呼び，α, β, γは各誤差の重み（ハイパーパラメータ），rは実画像，sは合成画像である．

$$\mathrm{L}_{total} = \alpha L_{GAN} + \beta L_{Aut} + \gamma L_{For}$$

Human Pose Estimation with Image Translation to Synthesized Image

Ryuhei TAKAHASHI †
Masaaki IIYAMA ‡
Motoharu SONOGASHIRA ‡
Atsushi HASHIMOTO ††
† Graduate School of Informatics, Kyoto University
‡ ACCMS, Kyoto University
†† OMROM SINICX Corporation

・敵対的誤差[1]

$$L_{GAN}(r,s) = |D_R(r)|^2 + \left|1 - D_R\left(G_R\left(E_S(s)\right)\right)\right|^2 + |D_S(s)|^2 + \left|1 - D_S\left(G_S\left(E_R(r)\right)\right)\right|^2$$

生成器により変換された画像を偽物の画像としてディスクリミネータに入力し，本物か偽物かを判定させ損失を計算する．なお，提案手法では，実画像デコーダG_Rの出力と実画像r，合成画像デコーダG_Sの出力と合成画像sをそれぞれ判別する2つのディスクリミネータを用いる．

・再構築誤差[2]

$$L_{Aut}(r,s) = \left|G_R\left(E_R(r)\right) - r\right| + \left|G_S\left(E_S(s)\right) - s\right|$$

敵対的誤差のみでは変換時に違う姿勢をとった人物画像を生成してしまう場合がある．再構築誤差を導入することで，画像を復元するために姿勢の情報を特徴量として抽出するようになり，生成器が姿勢を考慮した画像変換をするようになることが期待できる．なお，提案手法では，実画像→実画像の再構築誤差と合成画像→合成画像の再構築誤差の2つを用いる．

・変換誤差

$$L_{For}(r,s) = \left|G_R\left(E_S(s)\right) - s\right| + \left|G_s\left(E_R(r)\right) - r\right|$$

変換前の画像と変換後の画像の差をとったものである．再構築誤差と同様，姿勢が変わらないようにするために導入している．なお，この誤差の重みを大きくしすぎると，生成器が画像を全く変換しなくなるため，他の誤差よりも小さな重みを設定している．

なお，ネットワークに入力する前に，前処理として，実画像に対して背景差分を行って人物領域を抽出し，実画像，合成画像共に画像を同サイズにリサイズする．なお，店舗に設置された固定カメラなどでの利用の場合は背景画像が得られるため，容易に背景差分可能である．

提案手法では，Shottonら[3]の手法に倣い，人物を31部位に分け，画像の各画素に対して31部位のどれに対応するかのラベルを付与した部位ラベル画像を用意し，それをターゲットとして学習したものを姿勢推定器として用いる．姿勢推定器は，図3に示すように，深度画像が与えられたとき，各画素に対してその画素が人物のどの部位に属するかの尤度を算出し，尤度から関節位置を求める．関節位置を求める手順は，Shottonらの手法と同様のものを用いる．

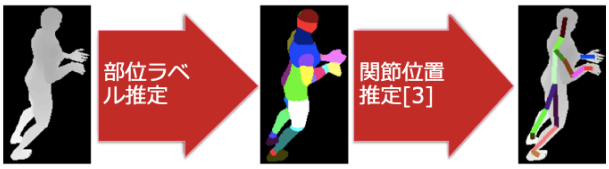

図3：姿勢推定器のフロー

3. 実験結果

以下に提案手法の適用例を示す．学習にはKinectで撮影した実画像および3DCGモデル生成ソフト「Poser」で生成した合成画像を256ピクセル×128ピクセルにリサイズしたものをそれぞれ4000枚用いた．図4に示すように，実画像に対してそのまま姿勢推定を行った場合は正しく推定できていないが，提案手法の画像変換を適用することで，人物のノイズや欠損が取り除かれ，正しく推定することが可能となっている．

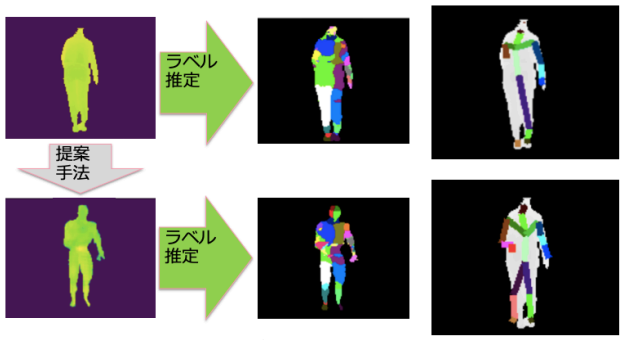

図4：提案手法の適用例

4. まとめ

本稿では，合成画像へ画像変換を行うことで，高コストである実画像のアノテーションをすることなく実画像に対して人物姿勢推定を行う手法を提案した．提案手法では人物を正面から写した画像に対してのみ適用可能である一方，実際に店舗内などでカメラが設置される場合，そのカメラの視点は多様である．今後，カメラの視点によらず画像変換が行えるように手法を改良し，想定している使用例に対しても適用可能にする予定である．

参考文献

[1] J.Zhu, T.Rark, P.Isola, A.Efros : “Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks”, ICCV(2017)

[2] A.Anoosheh, E.Agustsson, R.Timofte, L.Gool : ”ComboGAN: Unrestrained Scalability for Image Domain Translation”, CoRR(2018)

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