Transformerによる各関節の関係性に対する特注表現に着目した人体の2次元姿勢推定

小松悠斗† 平川翼† 山下隆義† 藤吉弘亘†
†中部大学
E-mail: u1370@mprg.cs.chubu.ac.jp

1 はじめに

人体の姿勢推定は，2次元画像上の人体の関節位置を推定する問題であり，モーションキャプチャや動作認識等に用いられる．これまでに人の姿勢変化に対応した手法が多数提案されているものの，特定の条件下では関節位置を正しく捉えられないことがある．例えば，一部の関節に対してオクルージョンが発生するシーンや，対象の関節と周囲の背景が類似するシーンが挙げられる．その原因として，対象の関節や関連する部位等の特徴情報をネットワークが正確に把握できていないことが挙げられる．本研究では，Transformerが人体の関節に対する大局的な関係性を捉えやすい特性に着目し，関節に対する多様な関係性を捉えた中間特徴を考慮した人体の2次元姿勢推定を提案する．これをマルチスケールなモデルとして構築することで関節に対する局所的な関係性から関節同士の大局的な関係性までを同時に考慮できる．また，ある関節とその関節に関連する部位に対してより着目させるためにAttention Convolution (Att-conv.) を提案する．

2 人体の2次元姿勢推定の従来手法

Deeppose [1] の登場以降，機械学習において広く研究されている人体の2次元姿勢推定は，DeepPoseのような関節位置を回帰により直接求める手法とConvolutional Pose Machine [2] のような各関節の位置をヒートマップとして出力する手法がある．各関節の位置をヒートマップとして出力する手法は，関節位置を回帰により直接求める手法に比べて，関節周辺だけでなく他の関節との関係性を捉えることが出来るため一般的となっている．

人体の2次元姿勢推定において，複数のスケールの特徴マップを利用することで，複数のスケールの特徴マップによる特徴表現を同時に考慮した人体の推定を行うことができる．そのためHourglass [3] など複数のスケールからなる特徴マップを考慮した研究が多数提案されている [4, 5, 6].

また，人体の姿勢推定にはTransformerを用いた手法も多く提案されている [7, 8, 9]．TransPose [7] は畳み込み処理で画像の特徴を捉えた後，連続したTransformer Encoderに順次入力することで人体の2次元姿勢推定を行う．これにより，畳み込み処理の手法に比べて周辺の関節だけでなく，より離れた関節との大局的な関係性を捉えることを可能としている．

3 提案手法

姿勢推定において，オクルージョンや背景との類似性に対応するためには関節の特徴を捉えるだけでなく，関節間の関係性やさらにそれらに関連する部位との関係性を捉えることが重要である．そこで，Transformerが畳み込み処理の手法に比べて人体の関節に対する大局的な関係性を捉えやすい特性に着目し，Transformerの中間特徴を利用した人体の2次元姿勢推定を提案する．本手法では，入力サイズが異なるTransformer Encoderから出力された中間特徴を集約することで，関節に対する局所的な関係から関節同士の大局的な関係の両者を考慮した推定ができる．また，対象の関節とその関連する部位に対してより着目する特徴を獲得するためにAttention Convolution (Att-conv.) を導入したAtt-conv. Transformer Encoderを提案する．

3.1 ネットワーク構造と中間特徴の利用

提案手法のネットワーク構造を図1に示す．提案手法は，事前学習済みのResNet [10] で構成されたバックボーンに画像を入力して，人体に対する特徴マップを求める．このとき，特徴マップのチャンネル数とサイズは $256 \times 32 \times 24$ となる．そして，特徴マップを平坦化した後，Att-conv. Transformer Encoder 1，2にて人体の局所的な特徴を捉える．その後，Overlapping Patched Embedding [11] により畳み込み層にて特徴マップを縮小し，Att-conv. Transformer Encoder 3に入力する．ここでは，関節に対する大局的な関係性を捉える．この処理を2回繰り返す．1回目の特徴マップのチャンネル数とサイズは $384 \times 16 \times 12$ となり，2回目では $512 \times 8 \times 6$ となる．縮小した特徴マップのサイズを拡大するために，1つの畳み込み層によって次元数を縮小前の2層のAtt-conv. Transformer Encoderと同様の次

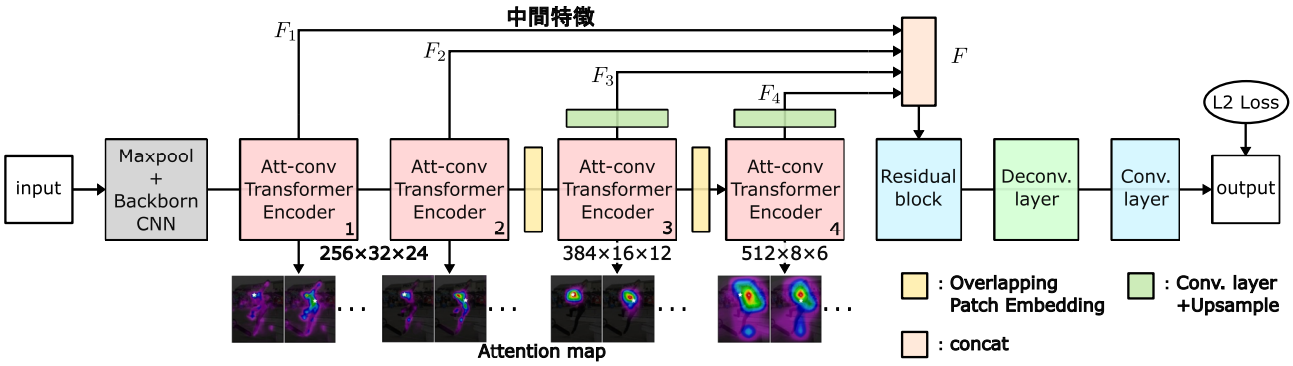

図 1　提案手法のネットワーク構造

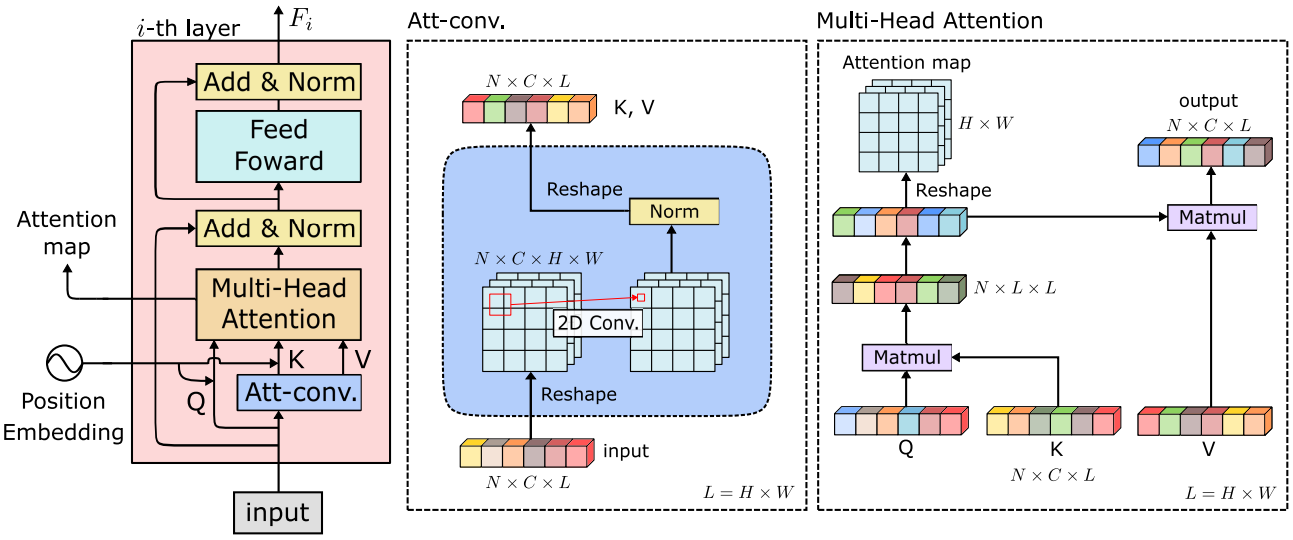

図 2　Attention Convolution Transformer Encoder の構造

元数に合わせた後，Upsampling する．4 層の Att-conv. Transformer Encoder の各特徴マップを式 (1) のように連結する．

$$F = F_1 \oplus F_2 \oplus F_3 \oplus F_4 \tag{1}$$

これにより局所的な関係と大局的な関係の両者を考慮した特徴マップとなる．その後，逆畳み込み層と畳み込み層による処理を施して各関節のヒートマップを出力する．

3.2　Att-conv. Transformer Encoder

Att-conv. Transformer Encoder の構造を図 2 に示す．Att-conv. Transformer Encoder は，Position Embedding の前に，畳み込み層で特徴マップを Key と Value に変換する Att-conv. を導入した Transformer Encoder である．これにより，Encoder に入力した特徴マップからより重要な領域の特徴を捉えた特徴マップを獲得できる．Att-conv. による処理の後，2D Sine Position Embedding による位置情報を Query と Key に埋め込み，それらを Multi-Head Attention に入力する．Query と Key の内積を Softmax により正規化することで，注視領域であるアテンションマップを獲得する．このアテンションマップと Value の内積 F' を求め，Multi-Head Attention の出力とする．

3.3　学習方法

提案手法では，ネットワーク最後の畳み込み層で獲得した推定ヒートマップと正解ヒートマップ間の誤差を L2 Loss により算出する．L2 Loss を式 (2) に示す．ここで，$\hat{y}_i$ は推定ヒートマップであり，y_i は正解ヒートマップである．

$$\mathcal{L} = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2 \tag{2}$$

4　評価実験

提案手法および提案手法において導入した各構造の有効性を示すために評価実験を行う．また，アテンションマップとヒートマップの可視化を行うことでネットワークが人体の関係性を捉えているか，また，中間特徴をもとに推定を行えているかを確認する．

4.1　実験概要

評価実験には，MS COCO データセット [12] を用いる．MS COCO データセットは，複数人の人々が写っている画像に対して関節位置でラベル付けされた 250,000 人のデータを含むデータセットである．学習時には，149,813

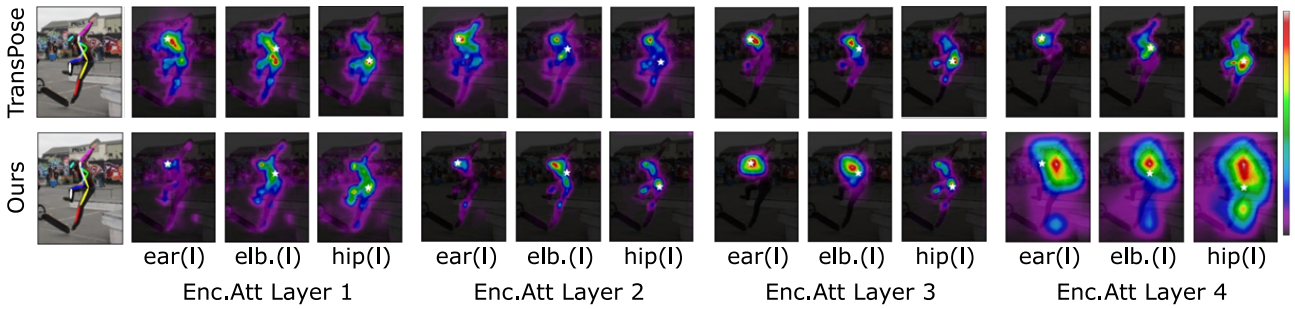

図 3　各 Encoder のアテンションマップの可視化例

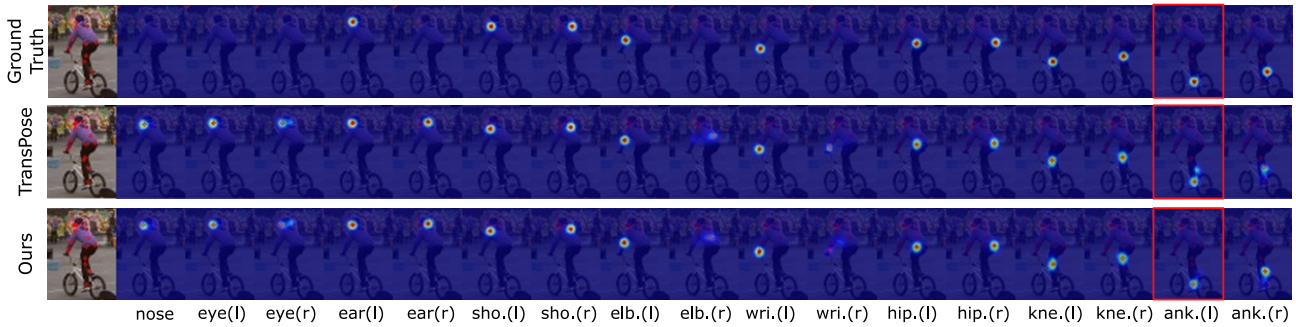

図 4　推定ヒートマップの可視化例

表 1　従来手法との評価比較

Method	# Params	GFLOPs	AP
Hourglass [3]	25.1M	14.3	66.9
CPN [4]	27.0M	6.2	68.6
SimpleBaseline [5]	68.6M	15.7	72.0
HRNet-W32 [6]	28.5M	7.1	73.4
HRNet-W48 [6]	63.6M	14.6	75.1
TransPose R-A4 [7]	13.4M	8.8	74.2
提案手法	24.0M	11.2	**75.5**

個のサンプルデータを用い，検証時には，6,325 個のサンプルデータを用いる．

提案手法はエポック数を 260，バッチサイズを 32 として学習する．最適化手法には，Adam を用いる．また，初期の学習率は 0.0001 であり，100，150，200，250 エポックで減衰させ，最終的に学習率 0.00001 で学習する．

実験を定性的に評価するにあたり，Object Keypoint Similarity (OKS) を利用した指標である Average Precision (AP) を用いる．OKS は，アノテーションされた関節における推定関節位置と正解関節位置の類似度の平均を表す指標であり，AP は，正解と推定したデータの中で実際に正解のデータである割合である．

4.2　従来手法との推定精度の比較

従来手法との評価比較では，AP の他にネットワークのパラメータ数と計算量である FLOPs (Floating point operations) を用いて比較を行う．従来手法との評価比較結果を表 1 に示す．提案手法は，Transformer をもとにした TransPose をベースにしているため，CNN を用いた従来手法と比べ，パラメータ数が少ない．また，従来手法と比べ，AP が向上していることが確認できる．しかしながら，スケールの縮小などのために畳み込み層を追加しているため，TransPose と比較するとパラメータ数が増加している．また，計算量である FLOPs も同様に増加していることが確認できる．

4.3　アテンションマップおよびヒートマップの可視化

各 Encoder のアテンションマップの可視化例を図 3 に示す．1 層目と 2 層目のアテンションマップでは，TransPose が人体の周辺全体に対してアテンションが発生しているのに対し，提案手法ではより人体の部位や関節にアテンションが発生していることが確認できる．提案手法の 3 層目のアテンションマップでは，人体の関節にアテンションが発生していることが確認できる．4 層目の TransPose のアテンションマップでは，部位や関節のまわりに局所的にアテンションが発生している．一方，提案手法のアテンションマップは，関節位置の周辺の他に，人物が唯一地面に接している左足にアテンションが発生していることが確認できる．これらのことから提案手法では，人体の関節に対する推定において重要な部分を画像中から捉えることができたといえる．

推定ヒートマップの可視化例を図 4 に示す．赤枠で囲んだ左足首の推定結果に注目すると，TransPose では，複数箇所にピークが発生している．それに対し，提案手法ではピークが一ケ所であることが確認でき，提案手法による改善を定性的に確認できる．

表 2　Att-conv. と中間特徴の利用による推定精度

Att-conv.	中間特徴	AP	AP^{50}	AP^{75}
		74.2	92.5	81.5
✓		74.5	92.5	81.5
	✓	75.2	92.5	82.6
✓	✓	**75.5**	**92.6**	**82.7**

4.4　Att-conv. と中間特徴の有無による推定精度

Att-conv. Transformer Encoder の導入と中間特徴を利用したネットワーク構造が，姿勢推定において有効であるかを実験により調査する．Att-conv. と中間特徴の利用方法による推定精度を表 2 に示す．Att-conv. を導入したネットワークは，導入していないネットワークと比較して AP が 0.3 pt 向上した．また，中間特徴を連結して利用したネットワークでは，AP が 1.0 pt 向上した．このことから，中間特徴を利用する方が Att-conv. を導入することよりも影響が大きいことが考えられる．また，両方の手法を組み合わせたネットワークでは AP が 1.3 pt 向上するため，両方の構造を導入することが最も有効であると確認できる．また，中間特徴を連結して利用したネットワークと両方の手法を組み合わせたネットワークは，AP^{50} では精度の変化は誤差程度だが，AP^{75} および AP において精度が向上しているため，正解関節位置により近づいた推定ができていると考えられる．

5　おわりに

本研究では，Transformer による中間特徴を考慮した人体の 2 次元姿勢推定を提案した．また，Att-conv. を Transformer Encoder に導入することで，対象の関節とその関連する部位に対してより着目する特徴を獲得することを可能とした．評価実験では，提案手法は従来手法を超える推定精度を達成し，また，導入した各構造においても導入前に比べ精度が向上しており，提案手法の有効性を確認した．今後の課題として，中間特徴に対する損失の適用などが挙げられる．

参考文献

[1] A. Toshev and C. Szegedy: “Deeppose: Human pose estimation via deep neural networks”, Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1653–1660 (2014).

[2] S.-E. Wei, V. Ramakrishna, T. Kanade and Y. Sheikh: “Convolutional pose machines”, Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pp. 4724–4732 (2016).

[3] A. Newell, K. Yang and J. Deng: “Stacked hourglass networks for human pose estimation”, European conference on computer visionSpringer, pp. 483–499 (2016).

[4] Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu and J. Sun: “Cascaded pyramid network for multi-person pose estimation”, Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 7103–7112 (2018).

[5] B. Xiao, H. Wu and Y. Wei: “Simple baselines for human pose estimation and tracking”, Proceedings of the European conference on computer vision (ECCV), pp. 466–481 (2018).

[6] K. Sun, B. Xiao, D. Liu and J. Wang: “Deep high-resolution representation learning for human pose estimation”, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5693–5703 (2019).

[7] S. Yang, Z. Quan, M. Nie and W. Yang: “Transpose: Towards explainable human pose estimation by transformer”, arXiv preprint arXiv:2012.14214 (2020).

[8] C. Zheng, S. Zhu, M. Mendieta, T. Yang, C. Chen and Z. Ding: “3d human pose estimation with spatial and temporal transformers”, Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2021).

[9] Y. He, R. Yan, K. Fragkiadaki and S.-I. Yu: “Epipolar transformers”, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7779–7788 (2020).

[10] K. He, X. Zhang, S. Ren and J. Sun: “Deep residual learning for image recognition”, Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778 (2016).

[11] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo and L. Shao: “Pyramid vision transformer: A versatile backbone for dense prediction without convolutions”, IEEE ICCV (2021).

[12] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár and C. L. Zitnick: “Microsoft coco: Common objects in context”, European conference on computer visionSpringer, pp. 740–755 (2014).