Transformer による中間特徴を考慮した人体姿勢推定に関する研究

TP20012 小松悠斗　　指導教授：山下隆義

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，ある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からである．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，Transformer による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

2 次元姿勢推定は Convolutional Pose Machines[1] の登場以降，各関節の位置をヒートマップとして出力する手法が一般的となっている．これら従来の手法は，複数の畳み込み層を積層することで，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ようになった．また，Transformer を用いた姿勢推定手法も多く提案されている．TransPose[2] 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，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ができている．

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 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．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[3] 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 回繰り返し，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．

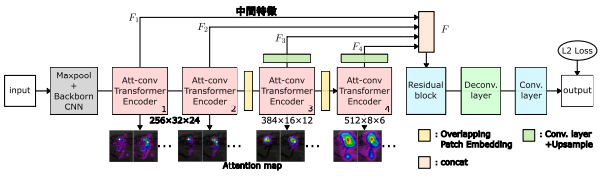

図 1: 提案手法のネットワーク構造

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

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．

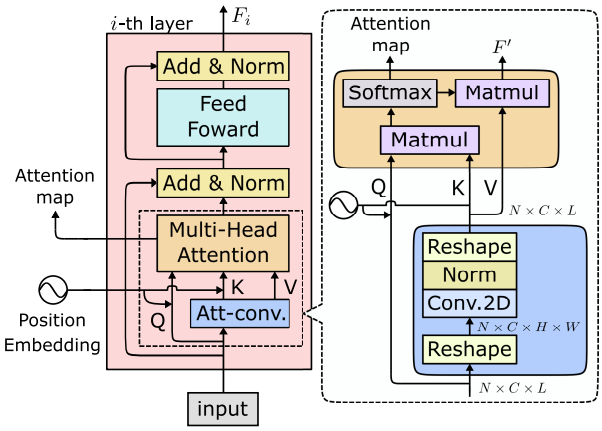

図 2: Att-conv. Transformer Encoder の構造

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．

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

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．

4. 評価実験

提案手法の有効性を確認するために，Transformer を用いた姿勢推定の従来手法である TransPose と精度比較を行う．データセットには，MS COCO データセットを用いる．評価方法は，推定姿勢と正解姿勢の適合率である Average Precision (AP) を用いる．

4.1 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

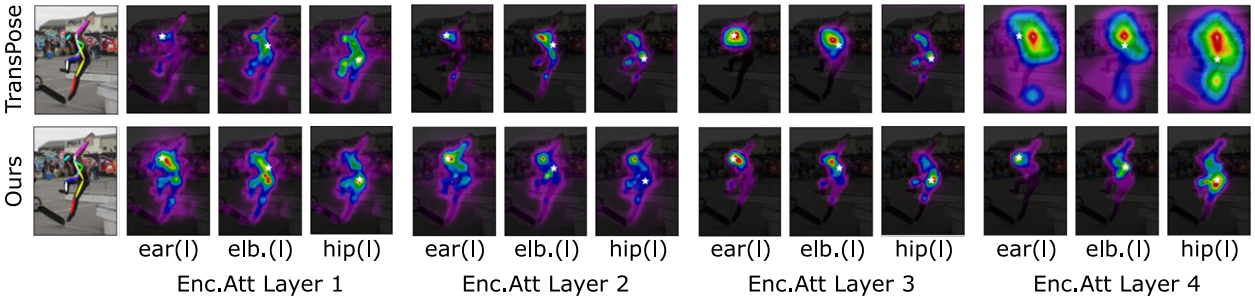

図 3: 各 Encoder のアテンションマップの可視化例

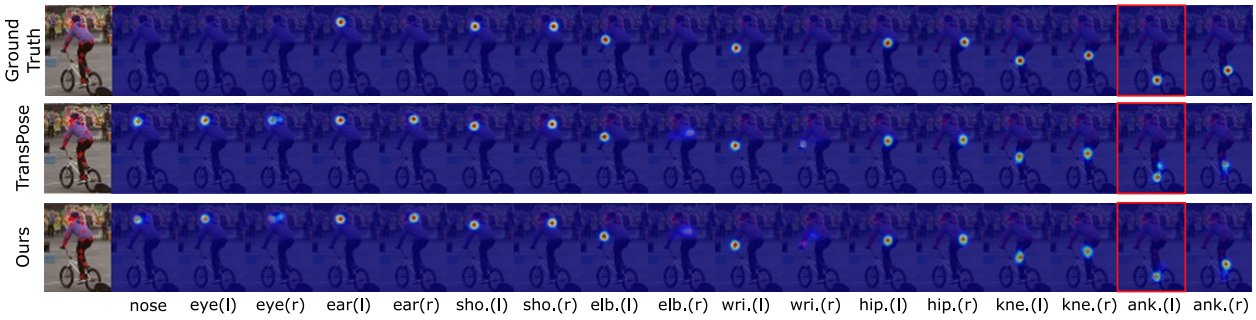

図 4: 推定ヒートマップの可視化例

表 1: 従来手法との評価比較

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

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．

4.2 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．

4.3 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

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**

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] S. Wei, *et al.*, “Convolutional Pose Machines”, CVPR, 2016.

[2] S. Yan, *et al.*, “Transpose: Towards explainable human pose estimation by transformer”, arXiv, 2020.

[3] W. wang, *et al.*, “Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction without Convolutions”, ICCV, 2021.

研究業績

[1] 小松悠斗 等, “2 次元空間での奥行きベクトル場を考慮した人物の 3 次元姿勢推定”, 第 27 回 画センシングシンポジウム, 2021.
(他 1 件)