技術論文

高精度及び高フレームレートな End－to－End多人数姿勢推定

戸部田　雅一
Masakazu Tobeta
鄭　澤
Ze Zheng
高椋　佐和
Sawa Takamuku
澤田　好秀
Yoshihide Sawada

概　要

人の行動を認識するために必要不可欠な技術である多人数姿勢推定を高フレームレートで実現するため,本稿では一般的なピラミッド特徴生成部と独自のヘッドからなる新たなモデルを提案する.このモデルを用いることで,他の手法で採用されている2段階推定や複雑な後処理を排除したアルゴリズムが実現できる.

COCOデータセットを用いた実験の結果,高い精度を維持しつつ,高いフレームレートを達成できたため報告する.

1.　はじめに

多人数姿勢推定は,群衆とロボットのインタラクションに不可欠な技術であると考えられており,モビリティ分野を含め,広く研究がなされている.しかし,これらの研究の多くは高い精度を実現するため,推論時にも大規模な計算機環境が必要となる[1)2)3)].そのため,リソースの限られた安価なエッジデバイスでリアルタイム処理が必要となる場合に,これらの研究手法を組み込むことは困難である.とくに車載ECUは,コストのみならず高い省電力性が求められるため,大規模な計算を必要とする多人数姿勢推定を車両へ搭載することは難しい.

多人数姿勢推定は大きく2つのアプローチに分類される.1つは物体検出器[4)]と関節点(例えば,鼻,左肩,右肩などの)尤度マップ推定器の2つのモデルを用いるトップダウン手法であり[3)],もう1つは関節点尤度マップと関節点間の関連度マップを一括推定するボトムアップ手法である[2)5)].

トップダウン手法では2つのモデルを用いるため,中間処理のオーバーヘッドと検出人数に応じた回数の関節点尤度推定が必要であり,ボトムアップ手法では関節点尤度マップと関節点間の関連度マップ統合のため複雑な後処理が必要となる.そのため,これら2つのアプローチはリアルタイム処理が困難であった.

本研究では,軽量な多人数姿勢推定の為の独自ヘッド,End-to-End Pose estimation Head (E2PH)を提案する.E2PHは軽量な畳み込みニューラルネットワーク(CNN)で構成され,2段階推定や高負荷な後処理を行わず複数人数の関節点座標値と尤度値を直接推定することが出来る.加えて,ResNet[6)]などのBackboneとFeature Pyramid Network (FPN)[7)]を用いることで,安価なエッジデバイスでも高い精度と高いフレームレートを維持できる.本モデルをEnd-to-End lightweight Pose estimation model (E2Pose)と呼ぶ.本稿ではまず関連研究について述べ,次に提案手法について説明する.

2.　関連研究

2.1　トップダウン手法

トップダウン手法では,物体検出器で生成したバウンディングボックス内に存在する1人の関節点尤度マップを推定する.これらの手法では,バウンディングボックスに基づく画像切り出しによって全ての人物をほぼ同じスケールに正規化でき,画像上の人物サイズばらつきに対して堅牢となる.そのため,様々な多人数姿勢推定ベンチマークにおいて高い精度を実現している[3)].しかし,検出人数に応じた回数の関節点尤度マップ推定が必要とされるため,人数が多くなるとフレームレートが低下する傾向がある[1)].

2.2　ボトムアップ手法

ボトムアップ手法では,画像内全ての人物に対する関節点尤度マップを推定し,このマップを人物毎にグループ化する.例えばOpenPose[1)]やPifPaf[5)]では関節点2点間を結ぶベクトルマップを用い,グループ化する.これらの手法では1回の推論演算で画像内全ての人物を検出できるため,特に人数が多い場合に高いフレームレートを実現できる.しかし実際には,正確な関節点を得るためには十分大きな関節点尤度マップと複雑な後処理が必要とされるため,全体の計算量が多くなる傾向がある.

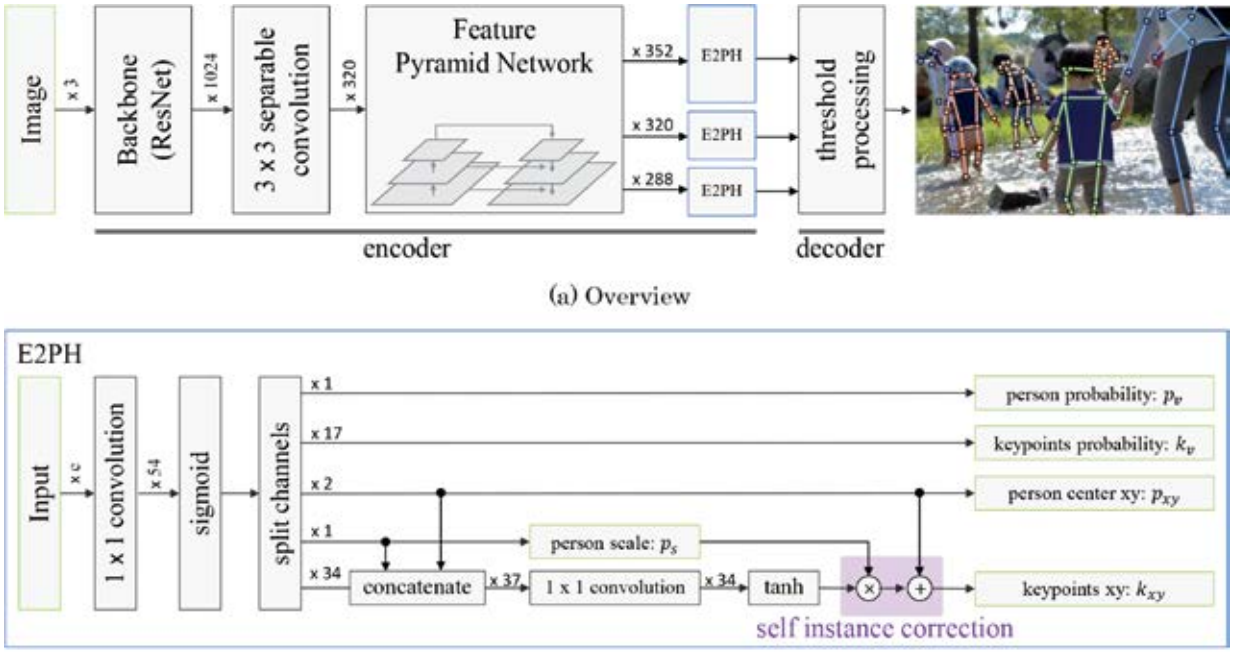

図1　(a): エンコーダは,Backbone,Separable convolution,FPN,E2PHブロックで構成され,デコーダはインスタンスの閾値処理を行う.(b): E2PHブロックは,次元変更,セルフインスタンス補正,活性化処理を行い,人物インスタンスを出力する.

2.3　End-to-end手法

End-to-end手法では,画像内全ての人物についてキーポイントを含む人物インスタンスを推定する.これにより後処理を大幅に削減できるが,先行研究のPOET[8]およびPETR[9]ではTransformer[10]を使用することに加え,検出人数に対するデコード演算を必要とするため,高フレームレートが実現できない.

3.　提案手法

本稿では,安価なエッジデバイスでリアルタイム実行可能な軽量な構成を用い,後処理を大幅に削減したEnd-to-end手法を提案する.図1(a)に我々のモデルの全体像を示し,以降で各ブロックの詳細を説明する.尚,本稿では人物インスタンスの正解をp,人物インスタンスの数をN_pとする.人物インスタンスは,人物存在尤度p_v,人物中心座標p_{xy},各関節点k=(k_v,k_{xy})から構成される.k_vは関節点の存在尤度,k_{xy}は関節点の座標とし,関節点数はN_kで表現する.なお,モデルの予測値は添え字ハットで表現し,例えば$\hat{p}$は予測された人物インスタンスを表す.

3.1　特徴抽出

提案手法では,画像から特徴マップを得るために3つのブロックを使用する.まず, ImageNet[11]によって事前学習されたResNet[6]をBackboneとして用いる.次にカーネル=(3,3)のSeparable Convolution,最後に高効率な多解像度特徴マップが得られるFPNを用いる.これらにより得られた特徴マップを後述するE2PHブロックへ供給する.

3.2　E2PHブロック

人物インスタンス$\hat{p}$を出力するため,図1(b)に示すE2PHブロックを用いる.E2PHブロックは特徴マップを入力とし,次元変更を行った後,POET[9]を参考に導入したセルフインスタンス補正と活性化処理を行う.

3.3　後処理

予測された全ての人物インスタンスに対して$\hat{p}_v \geq 0.5$と$\hat{k}_v \geq 0.5$の閾値処理を行い,無効なインスタンスを除却し,最終的な出力$\hat{\mathrm{p}}$を得る.即ち,他の手法で必要となる複雑な後処理は実施しない.

3.4　学習方法

E2Poseは固定数の人物インスタンス$\hat{\mathrm{p}} = \hat{p}_{(i)1:N_p} = \left(\hat{p}_{(i)xy}, \hat{k}_{(i,j)1:N_k}\right)_{1:N_p}$を出力する.ここで$\hat{p}_{(i)}$はi番目の人物,$\hat{k}_{(i,j)}$はi番目の人物のj番目の関節点を表す.学習の際には,$p_{(i)1:N_p}$と$\hat{p}_{(i)1:N_p}$の誤差行列を作成し,ハンガリー法[12]を用いた最小2部マッチングを求めることで,$p_{(i)}$と$\hat{p}_{(i)}$を1対1で紐づけて学習する.これにより,E2Poseは複数の人物インスタンスを重複なく出力できるようになる.なお,存在尤度p_vおよびk_vは誤差関数にFocal loss[13]を用い,座標p_{xy}およびk_{xy}は誤差関数に2乗誤差を用いる.

表1　COCO test-devセットに対する各モデルの性能比較.括弧内の精度は原著論文からの引用である.

Model	Detection image size	Pose image size	FPS ($\hat{N}_p$=4)	FPS ($\hat{N}_p$=13)	AP	AR
Top-down (the top 3 of speed):						
YOLOv3-D53 + ShuffleNetV2	320×237	256×192	13.3	5.3	49.0	57.5
YOLOv3-D53 + ShuffleNetV1	320×237	256×192	13.2	4.8	48.0	56.5
SSD-MobileNetV2 + ShuffleNetV2	320×320	256×192	12.1	5.1	45.2	52.4
Bottom-up:						
PifPaf (with the best AP)	None	512×512	5.4	4.1	58.6	61.8
PifPaf (with good AP/time ratio)	None	320×320	16.3	10.3	36.3	38.7
HigherHRNet-W32	None	512×512	0.6	0.3	66.4	71.3
End-to-end:						
POET-R50	None	512×512	6.3	6.1	(53.6)	(61.4)
(Ours)E2Pose-R50-512	None	512×512	19.5	19.5	54.3	60.5
(Ours)E2Pose-R101-512	None	512×512	18.9	19.2	59.6	65.1

4.　実験結果

4.1　実験設定

実験には,COCO keypoint estimation challenge（以降,単にCOCOと記す）のデータセットを用いた.COCOでは1画像あたり最大13人の人物インスタンスと,各人物に対する17個の関節点が設定されている.学習にはtrainセットを,検証にはvalidationセットとtestセットを用いた.BackboneはImageNetで事前学習された値で初期化し,その他のモジュールは乱数で初期化した.また,PifPafを参考に,画像の回転,切り抜き,左右反転,明るさ,コントラストのランダム水増しを行った.さらに,A. Bochkovskiyらにより提案されたMosaic augmentation[4]を適用した.入力画像サイズは512×512とし,その際に出力される人物インスタンス数はN_p=314とした.なお,これらのハイパーパラメータは実験によって適切に決定したものである.本稿では,多人数姿勢推定タスクにおいて標準的指標であるCOCO keypoint metricsのaverage precision（AP）を用いて評価した.評価の為のハイパーパラメータはCOCOで定義されている標準値を利用した.また,速度評価としてCOCO評価セットに含まれる全ての画像に対して,Jetson AGX XavierとTensorRT[14]を用いて推論時間を測定した.

4.2　実験結果

COCO validationセットの精度とフレームレートのベンチマーク結果を図2に示す.

この図は,提案手法を含む複数手法のネットワーク構造や入力画像サイズを様々に変更して作成しており,右上に位置するほど良いモデルであることを表している.この図から,E2Poseは精度を維持したまま高いフレームレートが実現できていることが分かる.これらの検証結果をもとに,速度面で優れたトップダウン手法3つ,精度面で優れたボトムアップ手法3つ,精度辺りの速度面で

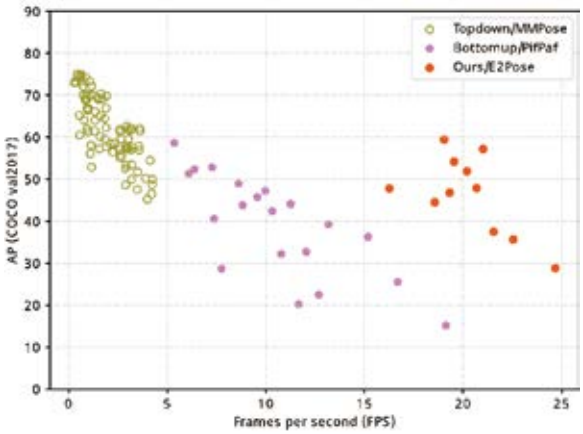

図2　精度(AP)と速度(FPS)のトレードオフ関係．FPSは検出人数枚に速度を平均し,その中央値を使用した.実験にはJetson AGX Xavierを用いた.

優れたボトムアップ手法,提案手法についてCOCO testセット（test-dev）を用い評価した.結果を表1に示す.APおよびARが認識精度,FPSは1画像につき4人検出（$\hat{N}_p$=4）された際と13人検出（$\hat{N}_p$=13）された際のフレームレートを示す.提案モデルは他の手法と比較し高い精度を維持したまま高いフレームレートを達成していることが確認できる.図3に提案モデルを用いた際の予測結果例を示す.

図3(a)は車両所有者の姿勢認識するために設定されたカメラ,図3(b)は車両周辺を監視するために設定されたカメラ,図3(c)は車室内を分析するために設定されたカメラをそれぞれ用いた予測結果である.提案モデルを用いることで,身体の一部が隠れている状況であっても人物姿勢を予測し,狭い空間で重なり合う人物も正確に推定できることが見て取れる.

(a) 車両所有者姿勢認識

(b) 車両周辺歩行者認識

(c) 車室内乗員認識

図3　車載カメラ画像に対する予測結果の例.モデルにはE2Pose-R101-512(COCO)を用いた.

5. おわりに

本稿では,リソースが限られたプラットフォームで高速かつ高精度な多人数姿勢推定を行うため,新しいEnd-to-end多人数姿勢推定フレームワークを提案した.軽量なE2PHブロックおよび後処理の大幅な削減により,COCOデータセットに対して高い精度(59.6AP)を達成しつつ,高いフレームレート(18.9FPS)を実現した.今後は,FPNやBackboneの改良による計算負荷の更なる削減と精度の向上を目指す.

本稿執筆にあたり社内外の関係者皆様より多くの協力が頂けたことに対し深く感謝いたします.

参考文献

1) Z. Cao, G. Hidalgo, T. Simon, S. Wei and Y. Sheikh: “Realtime multi-person 2d pose estimation using part affinity fields,” 2017
2) B. Cheng, B. Xiao, J. Wang, H. Shi, T. S. Huang, L. Zhang: “Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation,” 2020
3) R. Khirodkar, V. Chari, a. Agrawal, A. Tyagi: “Multi-Instance Pose Networks: Rethinking Top-Down Pose Estimation,” 2021
4) A. Bochkovskiy, C. Wang, H. M. Liao: “YOLOv4: Optimal Speed and Accuracy of Object Detection,” 2020
5) S. Kreiss, L. Bertoni, and A. Alahi: “PifPaf: Composite fields for human pose estimation,” 2019
6) K. He, X. Zhang, S. Ren, J. Sun: “Deep Residual Learning for Image Recognition,” 2016
7) T. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, S. Belongie: “Feature Pyramid Networks for Object Detection,” 2017
8) L. Stoffl, M. Vidal, A. Mathis: “End-to-end trainable multi-instance pose estimation with transformers,” 2021
9) S. Dahu, W. Xing, L. Liangqi, R. Ye, T. Wenming: “End-to-End Multi-Person Pose Estimation with Transformers,” 2022
10) A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin: “Attention Is All You Need,” 2017
11) J. Deng, W. Dong, R. Socher, L. Li, K. Li, L. Fei-Fei: “ImageNet: A Large-Scale Hierarchical Image Database,” 2009
12) H. W. Kuhn: “The Hungarian method for the assignment problem,” 1995
13) T. Lin, P. Goyal, R. Girshick, K. He, P. Dollar: “Focal loss for dense object detection,” 2017
14) H. Vanholder: “Efficient inference with tensorrt,” 2016

筆　者

戸部田　雅一
DS部
車両周辺監視関連の技術開発に従事

鄭　澤
DS部
MaaS/車室内監視関連の技術開発に従事

高椋　佐和
DS部
MaaS/車室内監視関連の技術開発に従事

澤田　好秀
DS部
人工知能に関する基礎・応用研究に従事