姿勢変化の連続性に着目した人物姿勢推定器の能動学習

森 太郎[1,a)]　出口 大輔[1]　川西 康友[2,1]　井手 一郎[1]　村瀬 洋[1]　井下 哲夫[3]

概要

深層学習に基づく人物姿勢の推定は，非常に高い性能が得られることから，近年様々な応用が検討されている．しかし，高い性能を得るためには大量の学習データが必要であることから，アノテーション作業の効率化への期待が高まっている．本発表では，ラベルなし時系列画像から人物姿勢推定器の学習に効果的な画像を自動で選択する能動学習手法を提案する．具体的には，人物姿勢は隣接フレーム間で連続的に変化する点に着目し，フレーム間での人物姿勢推定結果を比較することにより，誤推定の可能性がある少数の画像を能動的に選択し，人手によるアノテーション対象として抽出する．そして，人手によるアノテーションを施した少数のデータを追加することで人物姿勢推定器の性能向上を図る．実験を通じて評価を行ない，提案手法によって効果的な学習データ選択が可能なことを確認した．

1. はじめに

近年，深層学習に基づく人物姿勢の推定手法 [2], [3] が広く研究されており，その性能の高さから様々な形でその応用が検討されている．代表的な応用例として，人物姿勢を用いた行動認識などが盛んに研究されている [1]．しかし，高い性能を得るためには，画像中の人物の関節位置すべてを人手でアノテーションした学習データが大量に必要である．大量の画像に対して人物の関節位置を入力する作業は労力的・時間的に多くのコストを要するため，効率的に学習データを作成する技術が求められている．

効率的な学習データの作成手法として能動学習を利用した手法 [4], [5] がある．能動学習は，ラベルなし画像の集合から推定器の性能改善に寄与する画像を能動的に選択し，その画像に人手でアノテーションすることで，学習データを増強する手法である．学習に効果的な画像に絞ってアノテーションすることで，少ないコストで精度向上を実現することができる．

B. Liu らの手法 [5] では，人物姿勢の推定が不確かな画像をアノテーション対象として自動選択し，それらの画像に絞ってアノテーションすることで少ないコストで精度向上を実現している．この手法では，人物姿勢の推定結果ではなく，推定の前段階で作成するヒートマップを利用して不確かさを評価している．具体的には，ある人物の 1 つの関節のヒートマップに注目し，極大値が複数ある場合，つまり関節の候補が複数ある場合には，推定が不確かであると仮定している．そのためには，各関節のヒートマップを 1 人分ずつ推定する必要がある．したがって，人物姿勢推定手法のうち，人物検出を利用してそれぞれの人物に対して各関節のヒートマップを推定する top-down 手法 [2] に適用することはできるが，複数の人物に対して各関節のヒートマップを同時に推定する bottom-up 手法 [3] には適用できない．そのため，bottom-up な人物姿勢推定手法でも利用可能な能動学習手法が必要である．そこで本発表では，複数人物の関節位置を表すヒートマップに基づいて求めた人物姿勢推定結果を用いることにより，ラベルなし時系列画像から bottom-up 型の人物姿勢推定器の性能改善に寄与する画像を自動選択する能動学習手法を提案する．

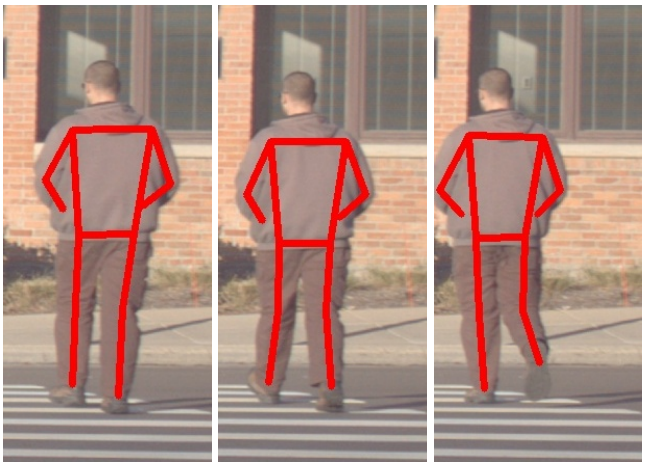
2. 提案手法

2.1 提案手法の概要

ある人物の姿勢を時系列に沿って観察すると，各関節の位置は連続的に変化することに気づく．図 1 に示すように，正しく人物姿勢を推定することができれば，隣接するフレーム間で推定される人物姿勢の差は小さい．逆に図 2 に示すように，人物姿勢の推定結果に誤りが含まれる場合，隣接するフレーム間で比較すると，中央のフレームと前後のフレームで人物姿勢推定結果が大きく異なることが考えられる．更に，前後のフレームと推定される関節の有無が異なる場合，いずれかのフレームで関節が未検出となっている可能性がある．本研究ではこのような知見に基づき，隣接するフレームで推定される姿勢が大きく異なる，もしくは隣接するフレームで推定される関節の有無が異なる場合には，誤推定の可能性がある画像と考え，能動学習におけるアノテーション対象として抽出する．

提案手法の処理手順を図 3 に示す．提案手法は，(1) 既存のラベルあり画像データセットを用いた初期人物姿勢推

[1] 名古屋大学
[2] 理化学研究所 情報統合本部 GRP
[3] 日本電気株式会社
[a)] morit@murase.is.i.nagoya-u.ac.jp



(a) $t-1$ フレーム　(b) t フレーム　(c) $t+1$ フレーム

図 1: 正しく推定できたときの姿勢変化．中央の推定姿勢はその前後フレームの推定姿勢と大きく異なっていない．

(a) $t-1$ フレーム　(b) t フレーム　(c) $t+1$ フレーム

図 2: 誤って推定したときの姿勢変化．中央の画像では左肘と左手首が誤推定されており，隣接フレームの左肘や左手首と大きく位置が異なっている．

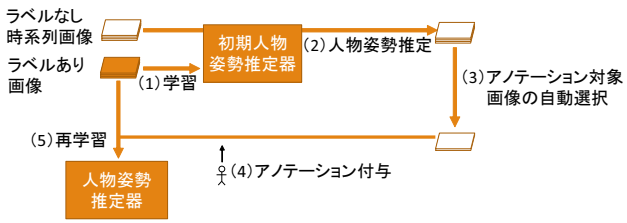


図 3: 提案手法の処理手順．

定器の学習，(2) 学習済み人物姿勢推定器を用いたラベルなし時系列画像からの人物姿勢推定，(3) 推定結果に基づいてアノテーション対象画像を自動選択，(4) 選択した画像への人手によるアノテーション付与，(5) 既存のラベルあり画像と追加アノテーション画像を用いた人物姿勢推定器の再学習，の 5 つの処理から構成される．以降，2.2 節で処理 (1)，(2) について説明した後，2.3 節で処理 (3) について，2.4 節で処理 (4)，(5) について詳しく説明する．

2.2　ラベルなし時系列画像からの人物姿勢推定

まず，基準となる初期人物姿勢推定器を構築する．ここでは，既存のラベルあり画像データセットで人物姿勢推定器を学習することで初期人物姿勢推定器を構築する．次に，構築した初期人物姿勢推定器を用いて，ラベルなし時系列画像に対して人物姿勢推定を行なう．その後，時系列画像中の各フレームにおける推定姿勢を対応付け，人物毎の推定姿勢系列を得る．ここで人物姿勢の対応付けには，人物姿勢追跡手法の PoseFlow[6] を使用する．

2.3　ラベルなし時系列画像から誤推定の可能性がある画像の自動選択

ラベルなし時系列画像から人物姿勢を誤推定している可能性がある画像を自動選択し，人手によるアノテーション対象画像として抽出する．具体的には，各画像で人物姿勢の誤り度を求め，求めた誤り度が高いものから順に，人手でアノテーションすべき対象として自動選択する．

ここで，$t \in \{1,...,T\}$ を時系列画像中のフレーム ID，$p \in \{1,...,P\}$ を人物 ID，関節 $j \in \{1,...,J\}$ の座標を $y^t_{p,j} = (u^t_{p,j}, v^t_{p,j})$ と表し，関節が推定されたか否かを $e^t_{p,j} \in \{0,1\}$ と表す．また，$e^t_{p,j} = 0$ のとき，$y^t_{p,j} = (0,0)$ とする．

まず，ラベルなし時系列画像中の各人物の大きさ S_p を推定姿勢から求める．複数画像に渡って存在する人物の推定姿勢は，未検出の関節の有無によって大きさが異なってしまう．そのため，ある人物 p の大きさを，時系列画像中で最も大きい人物姿勢を囲む矩形の面積と定義し，以下の式で求める．

$$S_p = \max_t \left\{ \left(\max_j(u^t_{p,j}) - \min_j(u^t_{p,j}) \right) \times \left(\max_j(v^t_{p,j}) - \min_j(v^t_{p,j}) \right) \right\} \quad (1)$$

次に，人物姿勢推定器から得られる人物姿勢を入力とし，ラベルなし時系列画像の各画像から誤り度を求める．時系列画像中のフレーム $t-1$，t，$t+1$ における推定姿勢を用いて，以下の手順によりフレーム t の誤り度を求める．

まず，推定された関節位置の違いに基づく誤り度を求める．フレーム t で推定された関節位置に誤りが含まれる場合，フレーム $t-1$ と t で推定された関節位置及びフレーム t と $t+1$ で推定された関節位置の変化が大きくなると考えられる．この考えに基づいて，関節位置の違いに基づく誤り度 C^t_L を求める．具体的には，フレーム $t-1$ と t 中の同一人物に対する推定姿勢における関節の Euclidean 距離の平均 $L^{t-1,t}_p$ 及びフレーム t と $t+1$ 中の同一人物に対する推定姿勢における関節の Euclidean 距離の平均 $L^{t,t+1}_p$

を以下の式で計算する．

$$L_p^{t-1,t} = \begin{cases} \frac{\sum_j e_{p,j}^{t-1} e_{p,j}^{t} \left\| y_{p,j}^{t-1} - y_{p,j}^{t} \right\|}{\sum_j e_{p,j}^{t-1} e_{p,j}^{t}} & \text{if} \quad \sum_j e_{p,j}^{t-1} e_{p,j}^{t} \neq 0 \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

$$L_p^{t,t+1} = \begin{cases} \frac{\sum_j e_{p,j}^{t} e_{p,j}^{t+1} \left\| y_{p,j}^{t} - y_{p,j}^{t+1} \right\|}{\sum_j e_{p,j}^{t} e_{p,j}^{t+1}} & \text{if} \quad \sum_j e_{p,j}^{t} e_{p,j}^{t+1} \neq 0 \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

フレーム $t-1$ と t の関節の距離及びフレーム t と $t+1$ の関節の距離の和 $\widehat{L}_p^t$ を以下の式で計算する．

$$\widehat{L}_p^t = L_p^{t-1,t} + L_p^{t,t+1} \tag{4}$$

求めた距離を各人物の大きさで重み付けする．すべての人物に対して上記の計算を行ない，総和をフレーム t の関節位置の違いによる誤り度 C_L^t として次のように計算する．

$$C_L^t = \sum_p S_p \widehat{L}_p^t \tag{5}$$

次に，関節が推定されたか否かに基づいて誤り度を求める．フレーム $t-1$ と $t+1$ で推定された関節がフレーム t で推定されていない場合，その関節は未検出であると考えられる．この考えに基づいて，未検出の関節数を数え，未検出による誤り度 C_U^t を以下のように求める．

$$C_U^t = \sum_{p,j} e_{p,j}^{t-1} e_{p,j}^{t+1} \left(1 - e_{p,j}^{t}\right) \tag{6}$$

最後に，求めた関節位置の違いによる誤り度 C_L^t と未検出による誤り度 C_U^t を正規化し，フレーム t に対する誤り度 C^t を以下の式で求める．

$$C^t = \frac{C_L^t - \min_t C_L^t}{\max_t C_L^t - \min_t C_L^t} C_L^t + \frac{C_U^t - \min_t C_U^t}{\max_t C_U^t - \min_t C_U^t} C_U^t \tag{7}$$

この処理をすべてのラベルなし時系列画像に適用し，それぞれの画像で誤り度を求める．

2.4 人物姿勢推定器の再学習

ラベルなし時系列画像からアノテーション対象画像を自動選択し，それらに対して人手でアノテーションを付与する．具体的には，ラベルなし時系列画像を入力として 2.3 節で求めた誤り度を降順に並べ替え，誤り度が大きい画像から順にアノテーション対象画像として選択する．そして，新しくアノテーションを付与した画像と既存のラベルあり画像を用いて人物姿勢推定器を再学習する．

3. 評価実験

提案手法の有効性を確認するため行なった評価実験について詳しく述べる．

表 1: データセットの内訳

学習		評価
ラベルあり	ラベルなし	ラベルあり
56,599 枚	3,188 枚	1,783 枚

3.1 データセット

評価実験では，Microsoft COCO データセット [7] と PedX データセット [8] の 2 つのデータセットを組み合わせて利用した．実験では PedX データセットでの評価を目的とし，まず異なる性質であるがデータ数の多い Microsoft COCO データセットで初期人物姿勢推定器を学習し，その後 PedX データセットのうち評価で利用しないデータを追加して人物姿勢推定器を再学習した．使用したデータセットの内訳を表 1 に示す．Microsoft COCO train set の 56,599 枚を学習用のラベルあり画像として利用した．PedX データセットは，ラベル付けされている画像と，ラベル付けされていない画像を含んだデータセットである．実験では PedX データセットのうちラベル付けされている画像を 2 つに分割し，3,188 枚をラベルなし時系列画像として学習に利用し，1,783 枚を評価用データとして利用した．ここで，PedX データセットは学習用・評価用で異なる画像を利用し，同一人物が含まれないように分割した．また，実験に利用した PedX データセットの学習用の画像は人物姿勢がアノテーションされているが，ラベルなし時系列画像として扱った．本実験では Microsoft COCO データセットと PedX データセットに共通してラベル付けされている関節を推定対象とし，nose, left eye, right eye, left shoulder, right shoulder, left elbow, right elbow, left wrist, right wrist, left hip, right hip, left knee, right knee, left ankle, right ankle の 15 関節を使用した．

3.2 実験方法

まず，学習用のラベルあり画像を用いて初期人物姿勢推定器を構築した．次に，初期人物姿勢推定器を用いて学習用のラベルなし時系列画像から人物姿勢を推定し，2.3 節に示す手順に従ってラベルなし時系列画像の各画像に対して誤り度を求めた．そして，求めた誤り度を利用してラベルなし時系列画像からアノテーション対象画像を自動選択した．なお本実験では，人手によるアノテーションを行なう代わりに，PedX データセットに付与されている人物姿勢を利用した．最後に，学習用のラベルあり画像に，新たにアノテーションを付与した画像を追加した学習データを用いて人物姿勢推定器を再学習し，評価用データによりその性能を調べた．

Bottom-up 型の人物姿勢推定手法である PifPaf [6] を対象とし，人手によるアノテーション対象画像の選択方法が異なる以下の 2 つの手法を比較した．

比較手法　ラベルなし時系列画像から無作為に画像を選

表 2: 各手法の人物姿勢推定精度（AP）

0%追加	20%追加（比較手法）	20%追加（提案手法）	100%追加
19.6	25.5	**26.2**	27.6

図 4: 提案手法によりアノテーション対象画像として選択した画像と，初期人物姿勢推定器による推定結果の例．

択して学習データとして追加とする手法

提案手法　ラベルなし時系列画像から誤り度が高い画像から順に画像を自動選択して学習データとして追加とする手法

評価指標には Average Precision（AP）を用いた．ここで，AP を計算するための推定姿勢の正誤判定には Object Keypoint Similarity（OKS）[7] を用いた．OKS は関節の推定値と真値の距離，人物の大きさ，関節ごとに決められた重みをもとに推定の正誤判定を決定する指標である．

3.3　実験結果

各手法について，ラベルなし時系列画像からアノテーション対象を 1 枚も選択しない場合（0% 追加），20% の画像を選択して追加して学習した場合，すべて追加して学習した場合（100% 追加），それぞれの精度を表 2 に示す．

4.　考察

表 2 より，隣接フレームにおいて推定される関節位置の違いと関節が推定されたか否かの違いを利用する提案手法は，無作為にアノテーション対象画像を選択する比較手法と比べて，人物姿勢推定精度を表す AP が 0.7 向上することを確認した．また，ラベルなし時系列画像の 3,188 枚すべてに対してアノテーションを付与した場合は AP が 8.0 向上するのに対して，提案手法ではラベルなし時系列画像から 20%（638 枚）の画像にアノテーション付与するだけで AP が 6.6 向上することを確認した．これは，3,188 枚の画像を学習データとして追加して再学習する際の精度向上量の約 8 割を，その 2 割のデータで達成できることを示している．このことから，提案手法では少数の学習データの追加で精度を大きく向上させることが可能なことを確認した．提案手法によりアノテーション対象画像として自動選択された画像と，それらに対する初期人物姿勢推定器による推定結果を図 4 に示す．提案手法では初期人物姿勢推定器では推定を間違える画像をアノテーション対象画像として選択できており，このような誤推定をする画像にアノテーションを付与して学習データとして追加することで，新たに姿勢を推定できる画像が増え，推定精度が向上したと考えられる．

5.　むすび

本発表では，ラベルなし時系列画像から人物姿勢推定器の性能改善に寄与する追加学習データを得るために，人手によるアノテーション対象画像を自動選択する能動学習手法を提案した．提案手法では，隣接フレーム間での推定姿勢の関節位置の違いと関節が推定されたか否かの違いを利用して，誤推定の可能性がある画像を自動選択し，アノテーション対象画像とした．評価実験により，無作為に選択する手法よりも高い精度が得られることを示した．

今後の課題として，時系列画像中の人物の動きの激しさの違いを考慮した手法の検討などが挙げられる．

謝辞 本研究の一部は科学研究費補助金 17H00745 による．

参考文献

[1] S. Yan, Y. Xiong, and D. Lin, “Spatial temporal graph convolutional networks for skeleton-based action recognition,” Proc. 23rd AAAI Conf. on Artificial Intelligence, pp.7444–7452, Feb. 2018.

[2] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, “Convolutional pose machines,” Proc. 2016 IEEE Conf. on Computer Vision and Pattern Recognition, pp.4724–4732, June 2016.

[3] S. Kreiss, L. Bertoni, and A. Alahi, “PifPaf: Composite fields for human pose estimation,” Proc. 2019 IEEE Conf. on Computer Vision and Pattern Recognition, pp.11977–11986, June 2019.

[4] D. D. Lewis and J. Catlett, “Heterogeneous uncertainty sampling for supervised learning,” Machine Learning Procs. 1994, pp.148–156, July 1994.

[5] B. Liu and V. Ferrari, “Active learning for human pose estimation,” Proc. 16th IEEE Int. Conf. on Computer Vision, pp.4363–4372, Oct. 2017.

[6] Y. Xiu, J. Li, H. Wang, Y. Fang, and C. Lu, “Pose flow: Efficient online pose tracking,” Proc. British Machine Vision Conf. 2018, no.53, pp.1–12, Sept. 2018.

[7] 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,” Proc. 13th European Conf. on Computer Vision, Part V, pp.740–755, Sept. 2014.

[8] W. Kim, M. S. Ramanagopal, C. Barto, M.-Y. Yu,K. Rosaen, N. Goumas, R. Vasudevan, and M. Johnson-Roberson, “PedX: Benchmark dataset for metric 3D pose estimation of pedestrians in complex urban intersections,” IEEE Robotics and Automation Letters, vol.4, no.2, pp.1940–1947, Jan. 2019.