2M-03

高速な物体検出手法を用いた動画からの効率的な人物姿勢推定

藤枝 慎　　山崎 俊彦　　相澤 清晴
東京大学

1. はじめに

画像中の人物の姿勢推定は、コンピュータビジョンの分野において古くから研究されてきた。その代表的な手法に、Pictorial structure model (PSM) [1, 2]やFlexible mixtures-of-parts model (FMM) [3]といった人物姿勢をパーツの連結によって表現するというものがある。

本研究では、カーネル密度推定を用いて、1フレーム中の複数の姿勢候補を統合するだけでなく動画の時間連続性を考慮して前後フレームの姿勢推定結果も統合することで、より高精度に人物の姿勢推定を行う方法を提案する。さらに、高速に一般物体検出を行うための手法である Faster R-CNN [4]を用いてフレーム中の人物位置の特定を行うことにより、処理の高速化を図る。

2. 関連研究

Yang ら [3]は HOG 特徴量 [5]を利用してパーツ毎に勾配方向に重みづけを行うことで様々な種類のテンプレートを作成し、異なる勾配をもつ一つの肢を複数のより小さなテンプレートの連結として表現することで、従来の手法を上回る成果を得ている。この Yang らが用いた手法を Flexible mixtures-of-parts model (FMM)と呼ぶ。しかし FMM では Double counting problem という、右脚と左脚のようによく似たパーツに対して誤ったテンプレートが適用されることにより同じ位置に二つの異なるパーツが推定されてしまうという問題が生じる。この問題のために、推定された姿勢が必ずしも最適であるとは限らない。

この問題に対して Cho ら [6]は、modified kernel density approximation (m-KDA)という手法を用いて、カーネル密度推定により推定スコアが上位 M 個の姿勢推定結果を統合するという解決策を提示している。

また近年、急速に発展している Deep Learning を用いた姿勢推定手法も登場している。Chen ら [7]は Deep Neural Network (DNN)を用いて、パーツの推定とそれらの位置関係を学習し姿勢推定を行う手法を提案している。

3. 提案手法

動画は連続した静止画の集合であり、この一つ一つの静止画のことをフレームと呼ぶ。動画の人体姿勢推定は一つ一つのフレームに対して、静止画における人体姿勢推定の手法を適用することで行うことができる。動画における各フレームは時間的に連続したものであり、フレーム間での時間変化も微小なものであると考えられるため、動画に関しては各フレームにおける姿勢は大きく変化するものではないと考えられる。この動画の時間的連続性に着目して、2章で述べた Cho らのカーネル密度推定を用いた手法を応用し、Figure 1 で示したように1フレームでの推定スコアにおける上位 M 個の推定結果に対してだけでなくその前後のフレームにおける推定結果に対しても密度推定を行うことで、姿勢推定のさらなる高精度化を図る。

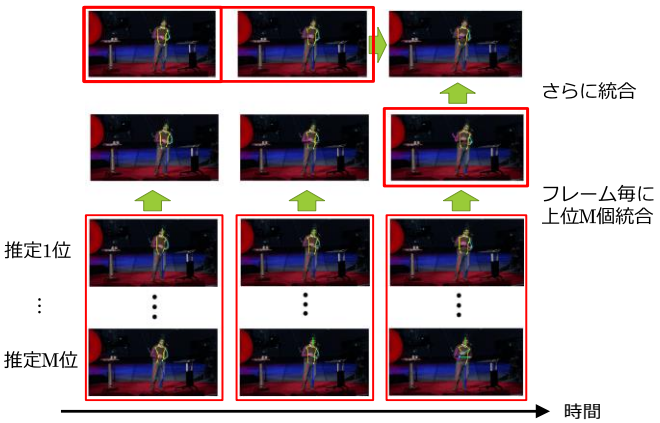

Figure 1　提案手法の概念図

さらにこの手法による姿勢推定を高速化するため、高速な物体検出手法である Faster R-CNN を用いる。提案手法において時間がかかる処理は、画像中の人物検出と HOG 特徴量から得られたテンプレートを用いてのパーツ推定である。そこでこのうち人物検出についての処理を、Faster R-CNN を使用することで高速化し、提案手法の処理時間の短縮を図る。

4. 実験

4.1 評価基準

本稿では実験結果を既存手法と比較する為、最も一般的な評価基準の1つである Percentage of Correct Parts (PCP) を用いる。しかし [3]でも述べられているように、PCP にはいくつかの異なる解釈がありそれにより結果が変わってしまう。そこで今回は strict PCP [7]という、ground truth と推定されたパーツの位置との間のずれが、ground truth におけるパーツの両端点間の長さに対して 50%以内であれば推定結果は正しいとする基準を用いる。なお ground truth はクラウドソーシングを利用することで作成した。

Efficient Human Pose Estimation in Video using Fast Object Detection.
Shin FUJIEDA, Toshihiko YAMASAKI, Kiyoharu AIZAWA, The University of Tokyo.

Figure 2　実験に用いた動画 (左から動画 A, B, C)

4.2 実験概要・結果

本実験では Figure 2 に示した 3 本の TED のプレゼンテーション動画[1]の一部を用いた。Figure 2 の左からそれぞれ動画 A, B, C とし、130, 340, 530 フレームの合計 1000 フレームである。2 章で言及した(a) Yang らの手法 [3]と(b) Cho らの手法 [6]、そして(c) 3 章で述べた提案手法を用いて人体姿勢を推定し、それらの精度を比較した。なお(b)と(c)においては推定スコアについて上位 16 個の推定結果を統合して最適な結果を得ている。さらに(c)では前後 2 フレームずつを含めた 5 フレームを用いて密度推定を行い、その重みは $w = \{0.85, 0.9, 1.0, 0.6, 0.4\}$ としている。

推定結果の strict PCP を Table 1, 2, 3 に示す。いずれの表も左の列から順に、使用した手法・胴・頭・上腕・前腕・大腿・下腿・平均を表している。Table 2, 3 を見ると、推定に失敗しているのはほとんどが腕であることがわかる。一方で Table 1 を見ると、腕だけでなく、下腿についても大きく推定に失敗していることがわかる。これは動画 A のスピーカーが途中大きく横に移動するために、正面を向いているフレームが他の動画と比べて少なく、脚が交差しているフレームが多いためだと考えられる。また Table 1 の L.arms と Table 3 の U.legs に関しては、(c)の提案手法が(b)よりも精度が低くなってしまっている。これは提案手法では前後フレームを含めて推定を行っているため、推定に大きく失敗したフレームが連続すると精度が落ちてしまうものと考えられる。上腕についての結果に注目するとその推定精度は、提案手法(c)は(a)に比べて平均で 8.2%、最新の手法である(b)と比べても平均で 6.5%向上していることがわかる。

次に、人物検出に faster R-CNN を用いることでどの程度処理が高速化したのかを確かめる。Figure 2 に示した 3 本の動画に対して、faster R-CNN を用いた場合とそうでない場合の提案手法の処理時間を比較した結果が Figure 3 で

Table 1　動画 A についての strict PCP

Method	Torso	Head	U.arms	L.arms	U.legs	L.legs	Mean
(a)	100.0	57.1	75.0	50.0	82.1	50.0	67.1
(b)	100.0	71.4	75.0	50.0	75.0	50.0	67.1
(c)	100.0	71.4	82.1	46.4	78.6	50.0	68.6

Table 2　動画 B についての strict PCP

Method	Torso	Head	U.arms	L.arms	U.legs	L.legs	Mean
(a)	100.0	96.3	81.5	50.9	100.0	99.1	85.9
(b)	100.0	96.3	85.2	51.9	100.0	100.0	87.0
(c)	100.0	96.3	88.9	55.6	100.0	100.0	88.5

Table 3　動画 C についての strict PCP

Method	Torso	Head	U.arms	L.arms	U.legs	L.legs	Mean
(a)	100.0	100.0	72.9	48.6	98.6	98.6	83.7
(b)	100.0	100.0	74.3	55.7	98.6	97.1	85.1
(c)	100.0	100.0	82.9	58.6	97.1	97.1	87.1

[1] TED ID はそれぞれ 1309, 2326, 1142

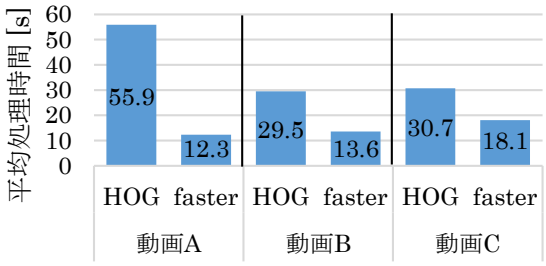

Figure 3　faster R-CNN の有無による処理時間の比較

ある。動画 A の処理速度が約 4.5 倍となっているのは、スピーカーが横を向いているフレームにおいては HOG 特徴量による探索が上手くいかず、1 フレームに約 400 秒かかってしまっているためである。動画 B, C を見てみるとその処理速度は平均で約 2 倍となっていることがわかる。

5. 結論

本稿では動画の時間的連続性を考慮した高精度な人体姿勢推定手法を提案し、最新の物体検出手法である faster R-CNN を組み合わせることで高速化も実現した。結果として提案手法は既存手法に比べて高精度に人物の姿勢推定を行うことができ、さらにその処理速度は約 2 倍となることを示した。今後の展望としては、パーツ推定に関しても faster R-CNN を用いて行い、さらに高精度に人物姿勢を推定できるように改良していきたいと考えている。

謝辞

本研究の一部は科学研究費助成事業(26700008)、および Microsoft IJARC core 10、人工知能研究振興財団、放送文化基金の支援を受けて行われた.

文　献

[1] M. A. Fischler and R. Elschlager: “The representation and matching of pictorial structures”, IEEE Transactions on Computers, vol. 100, no. 1, pp. 67–92 (1973)

[2] M. Andriluka, S. Roth, and B. Schiele: “Pictorial structures revisited: People detection and articulated pose estimation”, IEEE Conference on Computer Vision and Pattern Recognition pp. 1014–1021 (2009)

[3] Y. Yang and D. Ramanan: “Articulated human detection with flexible mixtures-of-parts”, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, no. 12, pp. 2878-2890 (2013)

[4] S. Ren, K. He, R. Girshick and J. Sun: “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks,” arXiv (2015)

[5] N. Dalal and B. Triggs: “Histograms of oriented gradients for human detection”, IEEE Conference on Computer Vision and Pattern Recognition (2005)

[6] E. Cho and D. Kim: “Accurate Human Pose Estimation by Aggregating Multiple Pose Hypotheses Using Modified Kernel Density Approximation”, IEEE Signal Processing Letters, VOL. 22, NO. 4, pp. 445-449 (2015)

[7] X. Chen and A. Yuille: “Articulated Pose Estimation by a Graphical Model with Image Dependent Pairwise Relations”, Advances in Neural Information Processing Systems, pp. 1736–1744 (2014)