**コンピュータビジョンの研究の近年の動向**

**3年　4組　47番　　藤井 悠太**

**1. はじめに**

　近年, さまざまな画像検出アルゴリズムが紹介されている[1]. 単一の画像の中から特定の物体を検出することはもちろん, その物体の表情や姿勢, 内面的な部分さえも推定することが可能になり, さらに研究が進められている[2]. 単一画像からの人物検出と姿勢推定が複雑になるにつれて, 動画での人物検出や姿勢推定も盛んになってきている. しかし, 単一画像に比べて動画での人物検出や姿勢推定は発達が遅く, 現在研究が盛んに行われている. 私は, 動画での人物検出と姿勢推定についての研究を調べ, 現在の研究状況についてまとめ, これからの展望と期待を述べる.

**2. Detect-and-Track: Efficient Pose Estimation in Videos**

## **2-1. 新規性**

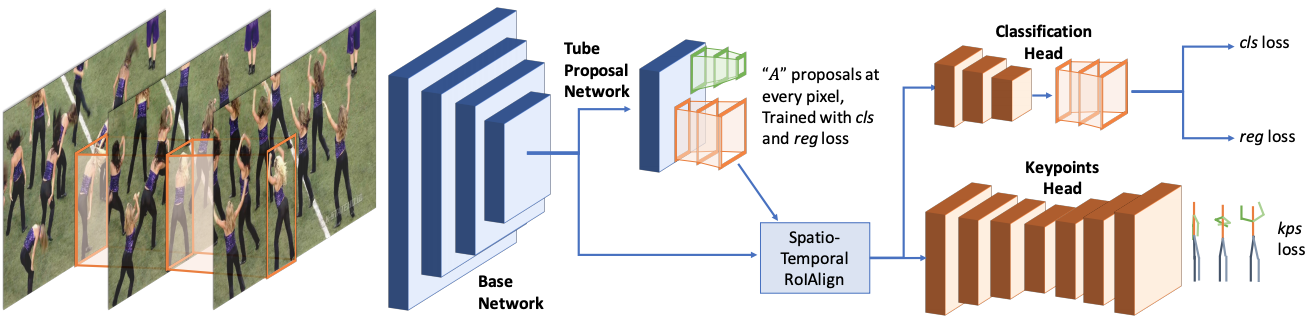
複数の人物が写り込んでいる単一の画像から特定の人物を検出するアルゴリズムについてはさまざまな研究がなされている. 有名なものの一つにYou Look Only Once(YOLO)がある[1]. 2016年に発表され, リアルタイムで検出処理を行うことができる高速なアルゴリズムである. 画像認識のアルゴリズムには他にも, R-CNNというアルゴリズムも有名である. R-CNNでは特徴量抽出とバウンディングボックスの検出で違うモジュールを

図1　Mask R-CNNのアルゴリズム（文献[3]より引用）

使用する. しかし, YOLOでは1つのネットワークで全てが完結するため, 比較的高速で高精度である. Detect-and-Track[3]では動画で姿勢推定を行うにはMask R-CNNというアルゴリズムを使うのが有効的ではないかと考えて研究を行なった. Mask R-CNNではYOLOやR-CNNのように検出した物体をバウンディングボックスで囲むだけでなく物体自体をピクセルレベルでマスクする(図2). この論文ではこれによって物体自体の形を捉え, 姿勢推定を行うことができるだろうと考えた. Mask R-CNNは本来複数の対象の中からある一つだけとは限らないパーツを特定し, 追跡するものである. Detect-and-Track[3]では, このMask R-CNNを時間軸方向に拡張し, 3次元にすることによって動画で姿勢推定を可能にした(図1).

**2-2. 有用性**

Detect-and-Track[3]では, PoseTrack dataset[5]という動画のデータセットを用いて畳み込みニューラルネットワークに通している. このデータセットには514の動画, 23,000のラベル付きフレームと, 1563,615のポーズが含まれている. その後, MOTAという精度を図るプラットフォームを用いて測ったmAP(mean Average Precision 平均精度)は図3のような結果になった. 平均精度が49.6とまだまだ精度は低い方だが, 動画での姿勢検出は可能であ

るということがわかった. またこの論文では, 画像で使用されるMask R-CNNのアルゴリズムを拡張しているので, 多くのGPUを必要とせず, そのため容量の大きい(800ピクセルまでの)動画を快適に読み込むことができると紹介している.



図2　○○○○○（文献[4]より引用）

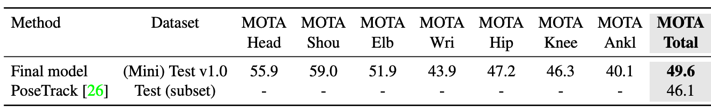


図3　○○○○○（文献[3]より引用）

**3. これからの展望と期待**

この先GPUの容量が増加し, このシステムが複数のGPUを横断してデータセットを学習することができるようになると, さらに高い平均精度を出すことができると考えている. Detect-and-Track[3]の論文内ではGPUの精度が向上したら高解像度, 大容量のデータセットを用いて実験を行いたいと述べられていた. 姿勢認識ができるとさまざまなインタラクションに発展するだろう. 例えば, 座った状態から立ち上がった時に視覚的にめまいを体験させるインタラクションなどである. Mask R-CNNを使用する場合は姿勢認識を行う対象者とは違う観覧者に対するインタラクションに有効的ではないだろうか. それらを現実的なものにするためには, Mask R-CNNの平均精度をさらに上げることが重要であろう.

**参考文献**

1. J. Redmon and A. Farhadi, “YOLO9000: Better, Faster, Stronger,” 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6517-6525, 2017.
2. Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, “Realtime multiperson 2d pose estimation using part affinity fields” 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.
3. R. Girdhar, G. Gkioxari, L. Torresani, M. Paruli, and D. Tran, “Detect-and-Track: Efficient Pose Estimation in Videos” 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018
4. Jonathan Hui, “Image segmentation with Mask R-CNN”, <https://medium.com/@jonathan_hui/image-segmentation-with-mask-r-cnn-ebe6d793272>, 2018.
5. U. Iqbal, A. Milan, M. Andriluka, E. Ensafutdinov, L. Pishchulin, J. Gall, and S. B. “PoseTrack dataset”, https://posetrack.net/about. php, 2017.