5R-04

YOLOv3 のアーキテクチャの改善によるリアルタイム Hand Gesture 認識及び家電操作への適用

小林　優太†　　ラシキア　城治†
中京大学†

1. はじめに

近年，様々なハンドジェスチャ検出（Hand Gesture Recognition（HGR））の研究が行われている．ジェスチャには静的なものと動的なものがある[1]. 静的 HGR は処理速度が速いが，限られたジェスチャしか扱えない[2][3]．動的 HGR は様々なジェスチャを対象にしているが，非常に自由度が高くジェスチャの開始や終了の判定が難しい．また，動的 HGR は 3DCNN を用いた検出[4]となるため処理速度が遅い．本研究では静的，動的ジェスチャのデメリットをカバーしたハイブリッド HGR を考える．ハイブリッド HGR とは静的ジェスチャを検出してから，検出した手の位置がどのように移動するかを追跡することで動的ジェスチャを認識する手法である．ハイブリッド HGR では手の位置を検出する必要があるため物体検出ネットワーク YOLOv3[5]を使用することにした．しかし YOLOv3 をそのまま使用して実験した結果，実行速度に問題があることが判明した．さらに Tiny-YOLOv3 も試みたが, 精度の向上が望ましかった．そこで YOLOv3 を新しいアーキテクチャに変更することを提案する．実験の結果から，提案手法のネットワークは Tiny-YOLOv3 と比べて約 2.6 倍高速化されており，精度の指標となる mean Average Precision（mAP）も高い水準を満たしている．また，提案ネットワークを実用的な家電操作に適用し，システムの有効性を確認できた．この論文で使用されるコードとトレーニング済みモデルは https://github.com/appleyuta/Hand-Gesture-Recognition に公開されている．

2. 提案手法

本研究では，実用的なシステムの開発を目的とし，実行速度などの問題を抱えている動的 HGR を避けハイブリッド HGR に注目した．ハイブリッド HGR とは静的ジェスチャを検出してから，検出した手の位置がどのように移動するかを追跡することで動的ジェスチャを認識する手法である．この手法を実装するには，物体の位置を検出する必要がある．そこで，物体検出技術の中でも安定した，処理速度の速い YOLOv3 を使用することにした. YOLOv3 は GPU などの高価なハードウェアで効率よく実行できるが，本研究では家庭にある通常の PC に備わっている CPU 上あるいはモバイルハードウェア（RaspberryPi など）でリアルタイムに実行できるようにネットワークの作成を目指した．

本研究では, まず YOLOv3 で HGR を行うためにネットワークを学習しシステムをテストした．しかし，推論の時間が遅くリアルタイムでのシステムの実現は難しいことが判明した．

YOLOv3 以外にも，高速な Tiny-YOLOv3 も開発されている．YOLOv3 と同じく Tiny-YOLOv3 でもシステムの構築をテストした．しかし，この方法では結果的に精度が落ちてしまいシステムで利用することは難しいと判明した．

そこで，YOLOv3 の Backbone を高速でかつ物体認識に有利な特徴量を抽出できるネットワークに変更することで速度と精度を両立したネットワークの作成を目的とした．

速度と精度を両立させたネットワークを構築することを目指して，様々なネットワークで Backbone を置き換えて実験を行った．実験結果により最終的に，MobileNetV3-Small [6]を Backbone に使用することになった．YOLOv3 の Backbone に MobileNetV3-Small を採用することで速度と精度の両立が可能になった．

Backbone と同様に Head 部分の高速化も検討した．Tiny-YOLOv3 の Head 部分は出力を 2 つのスケールのテンソルにすることで軽量化している．本研究では，Head の出力テンソルの数を削減することでネットワークを軽量化するのではなく，Convolution layer を Depthwise Separable Convolution に置き換えることで軽量化を目指した．

提案したアーキテクチャを MobileNetV3-Small YOLOv3 を省略して MNSY と呼ぶ.

ネットワークを学習するために Creative Senz3d Dataset を基にハンドジェスチャデータセットを集めた．クラスは 12 種類にし，画像の枚数は合計 33,000 枚を用意した．学習データに 8 割，検証データに 1 割，テストデータに 1 割を使用した．学習に Data Augmentation を用いた．

3. 応用例

本研究の最終目標はジェスチャで家電を操作するシステムを完成させることである．家電操作に必要となる機能を構築しアプリケーションを完成した．実装機能を以下に示し、システムの UI 構成を図 1 に示す.

① 操作対象の赤外線信号の登録・削除
② 赤外線信号とジェスチャの対応付け

Real-Time Hand Gesture Recognition Based on Improved YOLOv3 Architecture and Its Implementation for Home Appliance Control Systems
† Yuta Kobayashi, Lashkia George, Chukyo University.

Copyright ©2021 Information Processing Society of Japan.
All Rights Reserved.

③　HGR で家電の操作

今回はジェスチャ動作の種類として up，down，right，left の 4 種類を扱う．それぞれのジェスチャが認識されるとそれに対応する予め登録しておいた赤外線信号が送信される．家電操作の流れを図 2 に示す．

図 1 家電操作システムの UI 構成

図 2 家電操作の流れ

4. 実験結果と考察

本実験の実行環境は以下の通りである．OS は Windows10，CPU は Core i9-8950HK，ニューラルネットの作成に Tensorflow を使用した．更なる高速化を行うために，OpenVINO ライブラリも使用し Tensorflow のネットワークモデルから OpenVINO の IR モデルに変換を行った．OpenVINO はインテルが提供するさまざまなハードウェアでディープラーニング推論をより高速で実行するためのライブラリである．

物体検出の評価指標では，画像分類の時とは異なり物体の領域を検出することを含めて評価する必要がある．本論文では COCO Object Detection Challenge 評価指標で使用されている mAP を使用した．

得られた実験結果を表 1 に示す．提案した MNSY は Tiny-YOLOv3 に比べてパラメータ数，モデルサイズの両方で約 3.6 倍削減されている．実行速度を評価する指標として一般的に使われる fps(frames per second)を使用する．Tiny-YOLOv3 の 41.84fps に対して MNSY は 107.06fps であり，約 2.6 倍高速である．精度に関して，MNSY は Tiny-YOLOv3 に比べて mAP が 3.49%高い値を示した．MNSY はパラメータ数が少なく認識率は高いことが分かる．提案モデルの比較対象として Tiny-YOLOv3 の他に Google が物体検出に用いたモデル[6]MobileNetV3+SSDLite も使用した．これは MobileNetV3 を SSDLite の backbone に用いたモデルである．MobileNetV3+SSDLite は MNSY と比較してパラメータ数，モデルサイズが約 1.6 倍小さく，実行速度は約 1.8 倍高速であるが，精度の指標である mAP に関しては MNSY の方が 9.71%高い値を示した．MobileNetV3+SSDLite は実行速度が高速である一方，他のモデルと比較すると精度が著しく低下している．それに比べて，MNSY は速度と精度のバランスが良く様々な用途での応用が期待できる．

また，家電操作システムとしての有効性を確認するために，具体例としてテレビの操作を取り上げて実験した．ジェスチャの対応付けとしては，zero(握り拳)で電源管理，zero の左右動作によってチャンネル変更，zero の上下動作によって音量調整，チャンネルに対応する指の本数でチャンネル変更を行えるようにした．その結果，電源管理，チャンネル変更，音量調節をジェスチャによってスムーズに実行できることが確認できた．開発されたコード，トレーニング済みモデル及びデモ動画は

https://github.com/appleyuta/Hand-Gesture-Recognition

に公開されている．

表 1 他のネットワークと MNSY の性能比較

Model	Params	Model Size	mAP	Inference Speed
Tiny-YOLOv3	8.7M	33.2MB	81.26	41.84fps
MobileNetV3+ SSDLite	**1.5M**	**6.0MB**	75.04	**193.94fps**
MNSY	2.4M	9.2M	**84.75**	107.06fps

参考文献

[1] Sushmita Mitra, Tinku Acharya,“Gesture Recognition: A Survey”, IEEE Transactions on Systems, Man and Cybernetics- Part C:Applications and Reviews 37(3), 2007.
[2] Salem Ameen, Sunil Vadera,“A convolutional neural network to classify American Sign Language fingerspelling from depth and colourimages. ”Wiley Expert Systems, 2016.
[3] Adithya V., Rajesh R.,“A Deep Convolutional Neural Network Approach for Static Hand Gesture Recognition”, Procedia Computer Science, Elsevier, V. 171, 2353-2361, 2020.
[4] O. Köpüklü, A. Gunduz, N. Kose and G. Rigoll, “Real-time Hand Gesture Detection and Classification Using Convolutional Neural Networks”, IEEE International Conference on Automatic Face and Gesture Recognition, 2019.
[5] Redmon, Joseph and Farhadi, Ali “YOLOv3: An Incremental Improvement”, arXiv preprint arXiv:1804.02767, 2018.
[6] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang Vijay Vasudevan, Quoc V. Le, Hartwig Adam “Searching for MobileNetV3”, IEEE/CVF International Conference on Computer Vision, ICCV, 2019.

Copyright ©2021 Information Processing Society of Japan.
All Rights Reserved.