G-3

End-to-end モデルによる顔検出及び顔領域分割に関する検討

A Study on Face Detection and Face Segmentation using End-to-end Model

○古川　貴大[1]，関　弘翔[2]，細野　裕行[2]

*Takahiro Furukawa[1], Hiroto Seki[2], Hiroyuki Hosono[2]

Abstract: The purpose of this research is to create an End-to-end model for object detection and multi class segmentation, which are new tasks in image recognition. In this report, we study the creation of an End-to-end model for face detection and facial part segmentation.

１．まえがき

現在，様々な分野で深層学習が盛んに研究されており，特に画像認識分野において発展著しい．画像認識の主要タスクとして，物体識別，物体検出，セマンティックセグメンテーション（意味的領域分割）がある．さらに物体検出と領域分割を組み合わせ，検出した個々の物体毎に画素単位で領域を推定するインスタンスセグメンテーションまで実現さている．しかし，検出対象クラスとしての前景領域か背景領域か，すなわち 2 クラスを推定するにとどまっており，検出領域内のより詳細な意味的領域分割を実現するモデルは見受けられない．

そこで，本報告ではインスタンスセグメンテーションを実現する Mask R-CNN[1]を参考に，物体検出と検出物体のサブクラス，すなわち多クラスの領域分割を一貫して行うモデルに関する検討を行った．

２．報告内容

本報告では Mask R-CNN を基に，顔検出と顔器官ごとの領域分割（顔領域分割）を End-to-end で（一貫して）行うモデルに関して検討を行った．

Mask R-CNN とは，物体検出とインスタンスセグメンテーションを同時に行うマルチタスクのモデルである．Mask R-CNN は物体検出モデルである Faster R-CNN[2]を基に作成されている．Mask R-CNN が出力するセグメンテーションマスクは，検出対象クラスの前景領域か背景領域の２クラスである．しかし，例えば顔のように，顔クラスとしての前景か背景かだけでなく，目・鼻・口など，そのサブクラスのセグメンテーションマスクを推定できれば，より有益な情報になると考えた．これを実現するために，Mask R-CNN を改良し，多クラスに対応したマスクを出力するモデルを構築した．

提案モデルのモデル図を Fig.1 に示す．

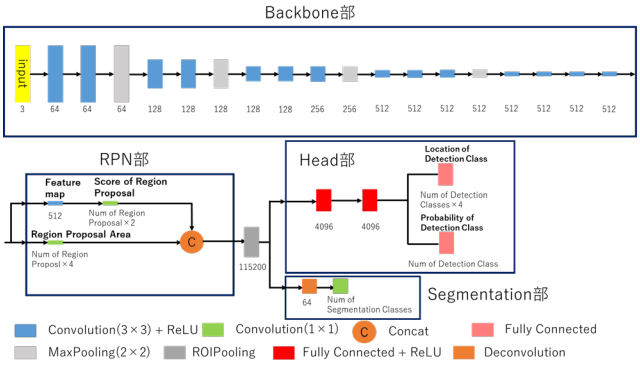

Figure 1. Our archtecture

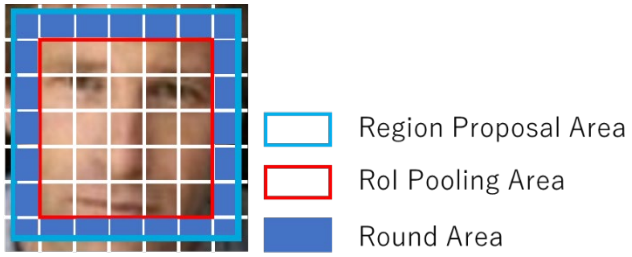

Figure 2. RoI pooling

２．１．Backbone 部

Backbone 部は VGG16[3]の構造を使用している．本報告では WIDER FACE DATASET[4]を学習した Faster R-CNN が内包する VGG16 から Fine-tuning を行う．Fine-tuning とは既存の学習済みモデルの重みを初期値として，新しいモデルを学習することである．ランダムに与えられた重みを初期値とするよりも，有効に顔の特徴が抽出できると考えた．

２．２．RPN 部

RPN(Region Proposal Network)部では，Backbone 部で抽出された特徴マップから物体らしき候補領域を抽出する．RPN 部で提案する様々なサイズの候補領域を RoI プーリングにより丸めながら同一の次元に揃えることで，Backbone 部で抽出した特徴マップをそのまま後段に流用でき，End-to-end 学習と高速化を実現している．RoI プーリングの概念図を Fig.2 に示す．

１：日大理工・院（前）・情報　２：日大理工・教員・情報

２．３．Head 部

Head 部では，RoI プーリングされた特徴マップに対して，全結合層により検出対象クラスの確率とその領域座標の推定を行う．

２．４．Segmentation 部

Segmentation 部では，RoI プーリングされた特徴マップに対して，検出対象の各サブクラス（本報告では，face, hair, eye, nose, mouth, hat, glass, background）に対応したセグメンテーションマスクを推定する．一般にセマンティックセグメンテーションのモデルはエンコーダ・デコーダ構造を取っている．そのため，提案モデルでは Backbone 部と RPN 部をエンコーダと捉え，Segmentation 部をデコーダのように逆畳み込みによるアップサンプリングと畳み込みにより構成した．

３．顔検出及び顔領域分割の実験

中部大学(MPRG)[5]にて配布されているデータセットを利用して提案モデルを学習し評価実験を行った．使用したデータセット内には 256×256 pixel の原画像と教師ラベルと顔領域の矩形座標が 13232 枚分含まれている．この内，10000 枚を学習用，1617 枚を学習時評価用, 1615 枚をテスト用とした. 顔検出では, WIDER FACE DATASET を学習した Faster R-CNN と提案モデルを，顔領域分割では，セマンティックセグメンテーション用の SegNet Basic を改良した顔領域分割モデル（SegNet Face）[6]と提案モデルを比較する．学習時のバッチサイズは 1, 最適化手法には Adam を採用した．使用した PC のスペックを以下に示す．

・GPU: NVIDIA TITAN RTX　・CPU: Intel Core i7 7700k

それぞれの評価指標として(1)式で与えられる IoU(Intersection over Union)を用いた．

$$\mathrm{IoU} = \frac{T_p}{G_t + F_p} \tag{1}$$

ここで，対象クラスの正推定領域をT_p，対象クラスの正解領域を $_t$，対象クラスの誤推定領域をF_pとする．また各クラスの IoU の平均値を mean IoU とする．テスト用画像に対する検出精度，領域推定精度，1 枚あたりの平均処理時間を Table 1 に，顔検出矩形と顔領域分割出力画像例を Fig.3 に示す．

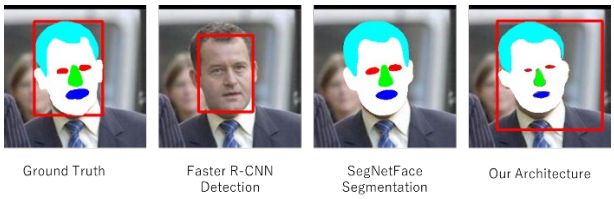

Figure 3. Result image

Table 1. IoU of face detection and facial part segmentation

	Detection			Segmentation	
	Faster R-CNN	Our Architecture		SegNet Face	Our Architecture
IoU	0.559	0.537	face	0.796	0.799
processing time	0.047	0.108	hair	0.669	0.551
			eye	0.513	0.196
			nose	0.646	0.606
			mouth	0.565	0.498
			hat	0.493	0.067
			glass	0.272	0.007
			background	0.949	0.400
			meanIoU	0.613	0.390
			processing time	0.026	0.108

Table 1 より顔検出の IoU は Faster R-CNN よりも低いことが分かる．顔領域分割では face 以外のクラスにおいて SegNet Face よりも低い IoU となった．また，Fig. 3 より目や口といった小さな顔器官に対する顔領域分割が正しく行えていないことが分かる．領域推定精度が低下した原因としては，RoI プーリングにより特徴マップの解像度が 15×15 まで低下していることが考えられる．

４．まとめ

本報告では，Mask R-CNN を基に，物体検出と多クラスの領域分割を End-to-end で行う新しいモデルの構築を, 顔を対象に検討した. 顔検出と顔領域分割を End-to-end で学習し，実行できたが，検出のために特徴マップを小さくすることから顔領域分割の推定精度が低くなるため，改善が必要である．

今後は，顔器官の中でも小さい器官である目や口の領域分割精度向上手法に関する検討を行う．

５．参考文献

[1] K. He, et al.: “Mask R-CNN,” IEEE ICCV2017 (2017)

[2] S. Ren, et al.: “Faster R-CNN:Towards Real-Time Object Detection with Region Proposal Networks,” IEEE PAMI 2017, Vol. 39, pp. 1137-1149 (2017)

[3] S. Liu, et al, “Very Deep Convolutional Neural Network Based Image Classification Using Small Training Sample Size,” The 3rd IAPR ACPR 2015 (2015)

[4] S. Yang, et al.: “WIDER FACE: A Face Detection Benchmark,” IEEE conf. on CVPR 2016 (2016)

[5] T. Yamashita, et al.: “Cost-Alleviative Learning for Deep Convolutional Neural Network-based Facial Part Labeling,” IPSJ trans. on CVA 2015, Vol. 7, pp. 99-103 (2015)

[6] 古川貴大,関弘翔,細野裕行：「SegNet Basic に基づく顔領域分割の精度向上に関する検討」,平成 31 年電気学会全国大会，3-112 (2019)