G-5

Semantic Segmentation を用いた顔領域分割
- SegNet Basic に基づく顔領域分割の精度向上に関する検討 -
Facial Part Segmentation using Semantic Segmentation
- A study on improvement of performance of facial part segmentation based on SegNet Basic -

○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 facial part segmentation using semantic segmentation is an important task to be the foundation of various researches such as facial expression recognition and face authentication. In this paper, we studied improvement of the performance of facial part segmentation based on SegNet Basic architecture.

１．まえがき

近年，深層学習の画像認識への応用例は多く存在し，発展著しい．画像認識における主要タスクとしては，物体識別，物体検出，Semantic Segmentation（意味的領域分割）がある．なかでも Semantic Segmentation は，画像内の各ピクセルに対してクラス識別を行うものであり，自動運転を始め様々な応用が期待されている．その身近な例の一つに顔領域分割がある．顔領域分割により顔器官の各領域を推定することは，表情認識や顔認証，機械読唇などの様々な応用研究の基礎となる重要な課題と考えられる．

本報告では，SegNet Basic[1]を改良し，メモリ増加量を抑えた顔領域分割の精度向上に関する検討を行った．

２．SegNet Basic

SegNet[1]とは畳み込みと Pooling による縮小を繰り返し行う Encoder と，Upsampling による拡大と畳み込みを繰り返し行う Decoder を有する Semantic Segmentation のモデルである．Pooling 時に選択した位置を Upsampling 時に参照して拡大することで，メモリ使用量を抑えつつ詳細な復元を実現する．SegNet Basic[1]は SegNet の軽量化モデルであり，畳み込みと Pooling を 4 つずつ持つ Encoder と Upsampling と畳み込みを 4 つずつ持つ Decoder で構成されている．

３．報告内容

SegNet Basic では Upsampling 時に Pooling で選択した位置を用いて値の復元を行うが，Pooling を繰り返し行うことで情報の損失が生じ，例えば領域境界などに関する詳細情報が欠落することが考えられる．この問題点に対し，Pooling の代わりに Dilated Convolution を用いて画像を縮小することなく広範囲の特徴を捉え，情報の損失を防ぐ手法を検討する．また，すべての Pooling を Dilated Convolution に置き換えた場合，特徴マップのサイズが大きいまま層を積み重ねることになるためメモリの使用量が増大し，SegNet Basic の利点を損なう．そこで，Encoder と Decoder の間を結合する Shortcut Connection と組み合わせる手法をあわせて検討した．提案モデルを Fig. 1 に示す．

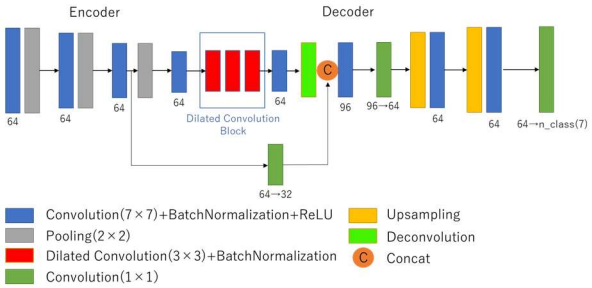


Figure 1. Our architecture

３．１．Dilated Convolution Block

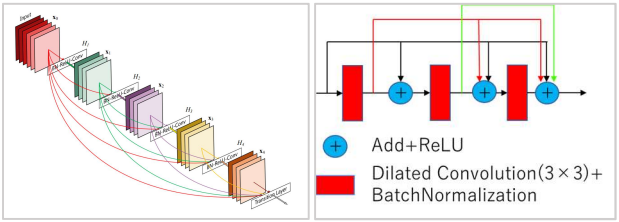
Dilated Convolution 層は，畳み込みの際に，フィルタの間隔（dilation）を変えて処理を行う層である[2]．通常は Convolution 層で畳み込みを行い局所的な情報を特徴マップとして抽出し，Pooling 層で集約して特徴マップが捉える空間領域を拡大する．しかし，Pooling 層を用いる場合，画像サイズが小さくなり，情報の損失が生じる．そのため Semantic Segmentation においては，領域の境界などに関する詳細な情報が欠ける問題が生じる．これに対し，dilation を大きくした Dilated Convolution 層を用いることで画像を極端に小さくすることなく情報を集約でき，広範囲の特徴を捉えることができる．本報告では dilation 2, 2, 4 の Dilated Convolution 層を積み重ねることで広範囲の特徴抽出を行う．このとき，局所的特徴と大域的特徴を同時に扱うために，Dilated Convolution 層間を結合する Dilated Convolution Block を検討する．複数の畳み込み層を結合する方法として有力なものに，Fig. 2(a)に示す Dense Block がある．Dense Block は，各層の出力チャネル数

１：日大理工・院（前）・情報　２：日大理工・教員・情報

を成長率として定め，その成長率分，次層の入力チャネル及び出力チャネルを増加させる手法である[3]．しかし，各層においてチャネルの変換や，結合を密に行うため，メモリ使用量が増加する問題がある．そこで Dilated Convolution 層を Fig. 2(b)に示すように，各層間の密な結合を同一チャネルの単純な加算とすることで，効率的な局所的特徴及び大域的特徴の抽出を提案する．

３．２．Shortcut Connection

Shortcut Connection とは任意の層間を結合し，特徴マップを加算あるいは連結する構造である．U-Net[4]では連結が採用され，Encoder の各 Pooling 前の情報を Decoder の Deconvolution 後の出力に結合している．これにより，Encoder 側の原画像に近い情報を Decoder 側で保持でき，復元性の向上が期待される．提案モデルでは，入力層に近い情報は解像度が高くメモリの使用量が増加に繋がるため使用せず，Encoder の終端に近い Pooling 前の情報を対応する Decoder へと連結する構造とした．その際に，Convolution(1×1)を用いたチャネル方向畳み込みによってチャネル数を半分にして結合することにより，省メモリ化を図る．



(a) Dense Block[3]　　(b) Dilated Convolution Block

Figure 2. Convolution block architecture.

４．顔領域分割の評価実験

中部大学(MRPG)[5]にて配布されているデータセットを利用して各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のペアが 13232 枚分含まれている．この内，10000 枚を学習用，1617 枚を学習時の評価用，1615 枚をテスト用とした．比較条件は，SegNet Basic，全ての Encoder と Decoder を Shortcut connection で連結するもの，Encoder と Decoder の間に Dense Block を配置したもの，及び提案モデルの 4 条件とし，学習時のバッチサイズは 24，最適化手法には Adam を使用した．

領域推定精度の確認指標として（1）式で与えられる IoU（Intersection over Union）を採用した．

$$IoU = \frac{T_p}{G_t + F_p} \qquad (1)$$

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G_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．

領域推定精度，及び学習時の VRAM（メモリ）使用量を Table 1 に示す．また，各条件での領域推定結果の例を Fig. 3 に示す．

Table 1. Accuracy of facial part segmentation

Value	SegNet Basic	Shortcut Conneciton	Dense Block	Our Architecture
face	0.7794	0.8091	0.7951	0.8034
hair	0.6554	0.7205	0.7071	0.7342
eye	0.4197	0.4163	0.4247	0.4248
nose	0.6235	0.6524	0.6426	0.6547
mouth	0.5892	0.6219	0.6041	0.6073
hat	0.3592	0.4778	0.4826	0.4982
background	0.9417	0.9592	0.9547	0.9593
mean IoU	0.624	0.6653	0.6587	0.6688
VRAM [GB]	6	7.8	7.2	7

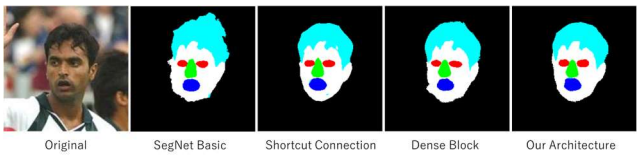


Figure 3. Result image of facial part segmentation

Table 1 より，各クラスにおける IoU は SegNet Basic に変更を加えることで向上することがわかる．ここで Fig. 3 より，Dense Block では境界領域の情報が損失し顎の輪郭が実際と異なるが，Shortcut Connection 及び提案モデルでは，実際の輪郭と遜色ない結果となることがわかる．また Table 1 より，Shortcut Connection では提案モデルよりもメモリの使用量が約 10%大きいことが分かる．以上より，メモリ使用量を抑えて領域推定精度を向上させる提案モデルが最も有効と考えられる．

５．まとめ

本報告では SegNet Basic の改良による，顔領域分割における領域推定精度の向上に関する検討を行った．Dilated Convolution Block や Shortcut Connection を効果的に用いることで既存手法の問題点である境界領域の情報損失を改善することができた．

今後は，MobileNet の採用などによるモデル軽量化と，精度向上を両立する手法に関する検討を行う．

６．参考文献

[1] V. Badrinarayanan, et al. :"SegNet :A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation.", IEEE trans. on PAMI, Vol. 39, pp. 2481-2495, 2017

[2] F. Yu, V. Koltun: "Multi-scale context aggregation by dilated convolutions", ICLR 2016 , arXiv:1511.07122, 2016

[3] G.Huang, et al. : "Densely Connected Convolutional Networks", IEEE conf. on CVPR 2017, pp. 2261-2269, 2017

[4] O. Ronneberger, et al. : "U-Net:Convolutional Networks for Biomedical Image Segmentation", MICCAI 2015, Vol. 9351, pp. 234 - 241, 2015

[5] 中部大学 MRPG　http://mprg.jp/　（2018 年 9 月現在）