H-024

Dice Loss と Multi Decoder Losses を用いた顔パーツのセマンティックセグメンテーション
Semantic Segmentation of Face Parts Using Dice Loss and Multi Decoder Loss

宮本 旭† 檜作 彰良† 中山 良平†
Asahi Miyamoto Akiyoshi Hizukuri Ryohei Nakayama

1. はじめに

顔の自動認識は，刑事裁判の被告人と防犯カメラに映った犯人の顔が同一であるかを専門家が判断する顔貌鑑定においても大きく期待されている．顔貌鑑定では，専門家が防犯カメラに映った犯人の顔のパーツ（目，鼻，口，顔の輪郭など）を特徴点として，そこから長さや距離，比率などの特徴量を手動で求め，鑑定結果を導く．顔パーツから求められた特徴量を用いることより鑑定精度は向上したが，手動で顔パーツを抽出する作業には多くの時間と労力を要し，その簡便化が切望されている．そこで，コンピュータにより，顔のパーツを自動抽出することができれば，鑑定時間の短縮と労力の低減が期待できる．

近年，畳み込みニューラルネットワーク（CNN: Convolutional Neural Network）を顔パーツのセマンティックセグメンテーションに応用した研究が報告されている．セマンティックセグメンテーションを目的としたネットワークの一つに SegNet[1]がある．SegNet は，Encoder でプーリングした位置座標を Decoder でアンプーリングする際，Pooling Index により空間情報を補っている．しかし，その情報が十分ではなく，顔パーツのセマンティックセグメンテーションに応用すると，小さなパーツの認識精度が低い問題があった[2]．そこで Aizawa らは，SegNet の Encoder から出力される異なる解像度の特徴マップを統合的に解析する Encoder-Multiple Decoders CNN（EMD）を提案し，異なる大きさの顔パーツを正確にセグメンテーション出来ることを報告した[2]．しかし，EMD は損失関数に交差エントロピー誤差を用いるため，データセットに含まれる各顔パーツの画素数が不均衡な場合，画素数が多いパーツの精度が高くなるように学習する傾向があった．

そこで本研究では，異なる解像度の Decoder 損失である Multi Decoder Losses と各顔パーツの画素数に応じて Dice Loss を重み付けした Weighted Dice Loss を組み合わせた新たな損失関数を定義する．また，スキップ接続を有することから，SegNet よりも空間情報の消失が少ないと考えられる U-Net をベースとした Encoder-Multi Decoders Network を構築する．そして，提案したネットワーク構造と損失関数による顔パーツのセマンティックセグメンテーション手法の有用性を評価する．

2. 提案手法

2.1 実験試料

実験試料として，Labeled Face in the Wild Dataset[3]で公開されている顔画像 13,233 枚（学習用：6,683 枚，評価用：6,550 枚）を用いた．これらの画像サイズは 250×250 画素であり，濃度分解能は 24bit（RGB color）である．これらの画像の背景／目／鼻／口／髪／眼鏡／帽子／それ以外の顔領域に対し，手動でラベルを付与した画像を学習時の教師データおよび評価時の正解データとして用いた．

† 立命館大学大学院 理工学研究科, Graduate School of Science and Engineering, Ritsumeikan University

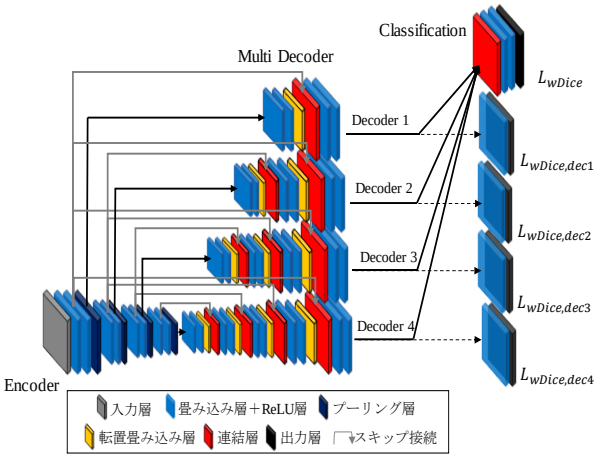


図 1　提案手法のネットワーク構造

2.2 提案手法のネットワーク構造

図 1 に提案手法のネットワーク構造を示す．提案ネットワークは U-Net をベースとした Encoder-Multiple Decoders Network であり，Encoder 構造，Multi Decoders 構造，Classification 構造をもつ．Encoder 構造は 8 つの畳み込み層と 4 つのプーリング層で構成され，Multi Decoders 構造は 10 の転置畳み込み層，32 の畳み込み層，10 の連結層および 4 つの出力層で構成される．また，Classification 構造は 1 つの連結層，2 つの畳み込み層，1 つの出力層で構成される．Encoder 側の 4 つのプーリング層のそれぞれから異なる解像度の Decoder 構造に分岐する．これら Decoder 構造で処理された特徴マップは，Classification 構造の連結層で統合される．最後に，統合した特徴マップに対して畳み込み処理を 2 回適用し，セマンティックセグメンテーションの結果画像を生成する．

また，Multi Decoders 構造の各 Decoder 構造の出力に畳み込み処理を 1 回適用し，各解像度の特徴マップに基づいたセマンティックセグメンテーションの結果画像を生成する．これらの結果画像は学習時の損失関数の計算で用いる．

2.3 提案手法の損失関数

提案手法の損失関数は，Classification 構造の Weighted Dice Loss L_{wDice} と Multi Decoders 構造の Multi Decoder Losses L_{MD} により定義した．

$$Loss = L_{wDice} + L_{MD} \quad (1)$$

Weighted Dice Loss は，正解ラベルとセマンティックセグメンテーション画像の重なりを評価し，下式で定義される．

$$L_{wDice} = 1 - \frac{2\sum_i w_i \sum_p A_{ip} B_{ip}}{\sum_i w_i \sum_p \left({A_{ip}}^2 + {B_{ip}}^2 \right)} \quad (2)$$

ここで，iは各クラス，pは画素数，A_{ip}は正解ラベル画像の One-hot ベクトル，B_{ip}はセマンティックセグメンテーション結果画像の One-hot ベクトルを表す．また，w は各クラスの重み係数を表し，下式で定義される．

$$w_i = \frac{1}{\left(\sum_p A_{ip}\right)^2} \tag{3}$$

L_{MD}は， 各 Decoder 構造の Weighted Dice Loss の平均を表し，下式で定義される．

$$L_{MD} = \frac{1}{M}\sum_M L_{wDice,decM} \tag{4}$$

ここで，MはEncoder構造から分岐したDecoder構造の数，$L_{wDice,\,decM}$はM番目のDecoderから出力されるWeighted Dice Loss を表す．

2.4 学習および予測

提案ネットワークの学習の最適化手法として Adam を用い，出力層の各画素の予測クラスと教師データのラベルの誤差が小さくなるように重みの更新を行う．また，ミニバッチサイズは 3，初期学習率は 0.001，エポック数は 10 とした．

2.5 評価方法

ネットワークの評価には，IoU（Intersection over Union）を用いる．IoU は正解ラベルとセグメンテーション画像の重複領域の画素数に基づき，下式で与えられる．

$$IoU_i = \frac{TP_i}{TP_i \cup FP_i \cup FN_i} \tag{5}$$

ここで，TP_iは正解ラベル画像とセマンティックセグメンテーション画像で各顔パーツクラスの領域 i が重なる部分の画素数を示す．また，FP_iはセマンティックセグメンテーション画像の領域 i が正解ラベル画像の領域 i と重ならない部分の画素数，FN_iは正解ラベル画像の領域 i がセマンティックセグメンテーション画像の領域 i と重ならない部分の画素数を表す．

3. 結果と考察

表 2 に，提案手法の各パーツに対する IoU の結果を示す．また，提案手法の有用性を評価するために，セマンティックセグメンテーションでよく用いられる SegNet，U-Net，そして EMD の IoU の結果も示す．提案手法のすべてのパーツの平均 IoU は 0.766 で，従来法の IoU（SegNet: 0.707，U-Net: 0.725，EMD: 0.752）より高い結果が得られた．各パーツの IoU では，鼻と口以外の 6 種のパーツで提案手法の IoU が最も高い結果であった．

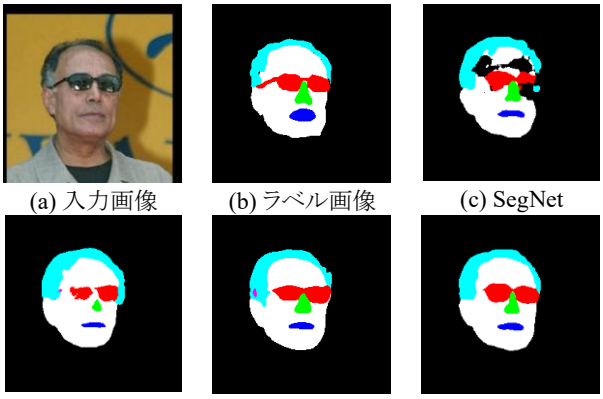
図 2 に各ネットワークによるセマンティックセグメンテーション結果画像を示す．提案手法は従来法と比較し，パーツの境界をより詳細にセグメンテーションできることが確認できた．提案ネットワークのスキップ接続により，適切に空間情報を補い，局所的特徴と全体的な位置情報を解析することができたこと，Multi Decoder Losses により細かな特徴も適切に学習させることができたことで，詳細な境界のセグメンテーションが可能になったと考える．

4. まとめ

本研究では，Weighted Dice Loss と Multi Decoder Losses を組み合わせた損失関数を持つ Encoder-Decoder 構造のネットワークにより，顔パーツのセマンティックセグメンテーションを行った． 提案手法のセマンティックセグメンテーションの精度は，EMD などの従来法よりも高く，その有用性が示唆された．

表 2　ネットワークの IoU 比較

	SegNet	U-Net	EMD	提案手法
背景	0.821	0.837	0.862	0.865
目	0.309	0.346	0.305	0.351
鼻	0.251	0.355	0.456	0.380
口	0.275	0.422	0.363	0.371
髪	0.564	0.564	0.607	0.613
眼鏡	0.131	0.131	0.129	0.140
帽子	0.136	0.173	0.146	0.182
顔領域	0.620	0.682	0.674	0.701
平均	0.707	0.725	0.752	0.766



(a) 入力画像　(b) ラベル画像　(c) SegNet

(d) U-Net　(e) EMD　(f) 提案手法

図 2　セマンティックセグメンテーション結果の比較

参考文献

[1] Badrinarayanan, Vijay, Ankur H., Roberto C., “Segnet: A deep convolutional encoder-decoder architecture for robust semantic pixel-wise labelling.”, arXiv:1505.07293 (2015)

[2] 相澤宏旭, 加藤邦人, 山下隆義,“Encoder-Multiple Decoders CNN によるセマンティックパーツセグメンテーション”, 電子情報通信学会論文誌 , Vol.102, No.6, pp.454-463 (2019).

[3] G. B. Huang, M. Ramesh, T. Berg, E. Learned-miller. “Labeled faces in the wild: A database for studying face recognition in unconstrained environments”, (2008)

[4] Ronneberger, Olaf, Philipp F.,Thomas B., “U-Net: Convolutional networks for biomedical image segmentation”., International Conference on Medical image computing and computer-assisted intervention (MICCAI), pp. 234-241 (2015).