G-012

舌の形状統一化のための **pix2pix** を用いた二値化舌画像の生成

Generation of binarized tongue image using pix2pix for shape unification

長谷川 豊† 綱島 克幸† 森 康久仁‡ 中口 俊哉§ 須鎗 弘樹‡
Yutaka Hasegawa Katsuyuki Tsunashima Yasukuni Mori Toshiya Nakaguchi Hiroki Suyari

1. はじめに

中国の伝統医学において舌は内臓の状態を映す鏡と言われている．その舌を診断に利用した舌診は，舌の色や形，または舌苔の色や厚さなどを観ることで身体の状態を診断するというものである．舌診の利点として即時性や非侵襲性であることが挙げられるが，定性的，主観的に基づいた診断を行うという性質のために限られた分野でしか用いられない．これを解決する手段として画像解析によって舌の特徴を定量化する研究[1]など，医工学として客観的に診断する研究が行われている．客観的に診断をする研究の一環として，我々は機械学習を利用することで，舌の表面の情報から舌診を行う試みを行っている．この際，舌の表面を学習する必要があり，学習時に舌の形状による影響をなくすために，舌の形状を標準舌形状に統一する必要がある．この統一する方法の一つに TPS(Thin Plate Spline)[2]と呼ばれる手法を用いたものがあるが，この手法では画像変形時の移動基準点として舌の重心を利用することから，重心を求めるために舌領域を白く，それ以外を黒くした二値化舌画像が必要となる．現状ではその作成が手作業で行われているため，研究の前処理に時間がかかる．そこで，本研究では効率よく研究を行えるようにするために，自動で二値化舌画像を作ることを目的としている．

二値化画像の代表的な生成手法として，閾値指定法や P タイル法，判別分析法といったものが挙げられる．これらの手法では画像の明るさや対象物の面積といったパラメータを元に二値化を行うが，画像ごとに最適なパラメータを設定する必要があり，そのパラメータを見つけることにも手間がかかる．また，これらの手法で生成した二値化舌画像は，舌領域を白く，それ以外を黒くしたものという目的の画像とは異なっている．そこで，今回必要としている画像は舌とそれ以外とのセグメンテーションと考え，現在様々な分野において活躍しているディープラーニングを利用したセグメンテーションに注目した．ディープラーニングは生成したい画像を学習させるだけで，多くの画像に対して学習させた画像のような画像を自動で生成でき，閾値指定法や P タイル法のように画像ごとに異なるパラメータを設定する必要がない．また，MRI 画像からの脳腫瘍のセグメンテーション[3]など高精度が求められている医療画像のセグメンテーションにディープラーニングが多く活用されており，舌画像に対しても精度のよいセグメンテーションを行い，手作業で作成したものと遜色のない二値化舌画像が生成可能なのではないかと考えた．今回の二値化舌画像の生成には多種多様な画像生成に用いられた実績のある pix2pix と呼ばれるアルゴリズムを利用する．

† 千葉大学大学院融合理工学府，Graduate School of Science and Engineering，Chiba University
‡ 千葉大学大学院工学研究院，Graduate School of Engineering，Chiba University
§ 千葉大学フロンティア医工学センター，Center for Frontier Medical Engineering，Chiba University

2. pix2pix

pix2pix[4]は画像生成アルゴリズムの一種であり，２つのペアの画像から画像間の変換方法を学習し，学習結果を元に入力された画像に対応する画像を生成するというものである．本節では pix2pix の特徴やその構成について述べる．

2.1 pix2pix の特徴

一般的に画像生成を行うには，特定の問題ごとにネットワークを設計する必要がある．しかし，pix2pix は汎用性の高い画像生成アルゴリズムであり，問題ごとにネットワークを設計する必要がない．よって，容易に画像生成を行うことができ，航空写真から地図の作成，白黒画像からカラー画像の生成，昼の風景から夜の風景の生成など多種多様な画像生成に用いられている．これには GAN と呼ばれる技術が用いられている．

2.2 GAN

GAN(Generative Adversarial Network)は訓練データを学習し，それらと似たような新しいデータを生成する生成モデルと呼ばれるネットワークである[5]．GAN は図 1 のような Generator と Discriminator の２つのネットワークからなり，Generator は訓練データと同じようなデータを生成しようとし，Discriminator はデータが訓練データであるか Generator によって生成されたデータであるかを識別しようとする．Generator は生成した画像を Discriminator によって生成した画像と見抜かれないように学習し，Discriminator は訓練データと生成されたデータを誤識別しないように学習していくことで，最終的に訓練データと同じような画像が生成される．このとき Discriminator の識別率は 50%であることが期待されている．pix2pix は Generator には U-NET，Discriminator には PatchGAN を用いている．

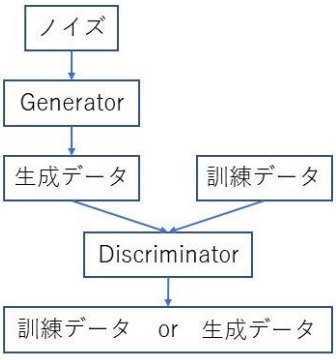

図 1 GAN の概要

pix2pix は GAN の中でも特に cGAN（conditional GAN）と呼ばれるものの一種である．GAN は Generator の入力値としてランダム値を利用していたが，cGAN は Generator と Discriminator に意味のあるノイズや画像，文字など生成の補助となる条件を加えることで，生成の精度を向上させるものである．pix2pix では画像を条件として使用している．

cGAN の目的関数は次のように表現される．

$$L_{cGAN}(G,D) = \mathbb{E}_{x\sim Pdata(x,y)}[\log \mathrm{D}(x,y)] + \mathbb{E}_{x\sim Pdata(y),z\sim p_z(z)}[\log(1-\mathrm{D}(\mathrm{G}(x,z)))]$$

$\mathrm{D}(x,y)$は訓練データを訓練データと判断した確率，$\mathrm{D}(\mathrm{G}(x,z))$は生成された画像を訓練データと判断した確率である．Discriminator はこの目的関数を最大化しようとするのに対し，Generator は最小化しようとする．

Generator は Discriminator をだますような画像を生成するだけでなく，正解により近い画像を生成する必要がある．これには cGAN の目的関数に以下の L1 ノルムを加えるのが効果的である．

$$L_{L_1}(G) = \mathbb{E}_{x,y\sim Pdata(x,y),z\sim p_z(z)}\left[\left|\left|y-G(x,z)\right|\right|_1\right]$$

L1 ノルムのみによる画像生成では細部がぼやけた画像が生成されるが，全体像を捉えることができる．一方，cGAN のみによる画像生成では詳細を捉えることができる．よってこれらを組み合わせることで精度の高い画像を生成できる．したがって，pix2pix の目的関数は以下のようになる．

$$\mathrm{G}^* = \arg\min_G \max_D L_{cGAN}(G,D) + \lambda L_{L1}(G)$$

2.3 U-NET

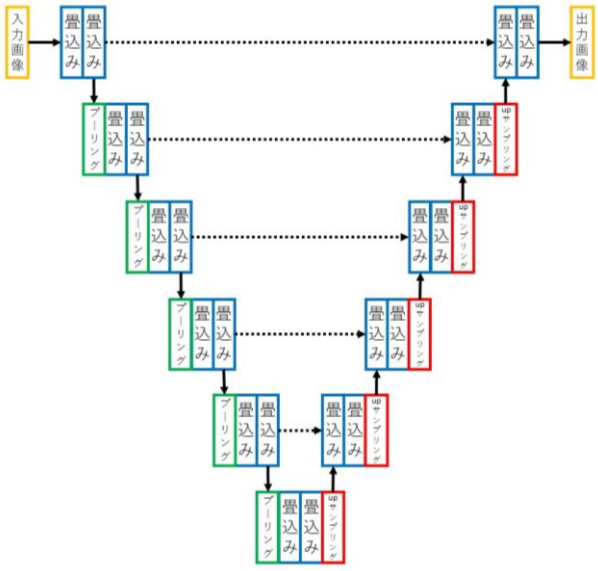

図 2 U-NET の構成図

CNN(convolutional neural network)によるクラス分類において，畳み込み層は画像内の物体の局所的な特徴を抽出し，プーリング層は画像内における物体の位置情報をぼかすことで，物体の位置ズレや大きさによる影響を小さくするという役割を担っている．そのため，層が深くなるほど，局所的な特徴を抽出することができ，その位置情報は曖昧になるため，精度の高い認識を行うことができる．しかし，領域抽出においては物体の局所的な特徴とその位置情報を元画像上で正確に復元する必要がある．そこで，局所的な特徴と位置情報の両方を学習できるように設計されたネットワークが U-NET[6]である．U-NET は図 2 のような Encorder-Decorder 構造をしている．Encorder 部では畳み込み層とプーリング層によって特徴を抽出する．層が深いほど抽出される特徴は局所的で，その位置情報は曖昧になり，層が浅いほど抽出される特徴は全体的で，その位置情報は正確になる．Decorder 部では畳み込み層とアップサンプリング層によって，特徴を保ったまま画像サイズを大きく復元できる．この時，Encorder と Decorder の対応した画像サイズの特徴を合わせることで，Decorder に位置情報を伝えることができ，局所的な特徴を保ったまま，位置情報を復元し，より詳細に画像を復元することができる．

2.4 PatchGAN

PatchGAN は，より詳細な画像を生成するための仕組みである[4]．Discriminator に入力された画像が訓練データであるか，それとも Generator によって生成された画像であるかを判断させる際に，画像全てではなく，16×16 や 70×70 といった小さな領域に区切った画像から判断させる．これにより，画像全体を見るのではなく，局所的な領域を見て判断するため，画像の高周波成分，つまり詳細な部分についての妥当性が得られる．一方，画像の低周波成分についての妥当性は L1 ノルムによって得られる．また，PatchGAN は小さな領域に区切ることからパラメータ数を減らすことができ，より効率的に学習できる．

3. 提案手法

(a)舌画像　　(b)二値化した舌画像

図 3 訓練画像の例

舌画像とそれを二値化した二値化舌画像のペアを学習データとして pix2pix で学習させ，任意の舌画像を入力した際に対応した二値化舌画像を生成できるようにする．

学習データとして同一環境下で撮影された 39 人分の被験者の舌画像を用意し，各被験者につき 6〜21 枚の舌画像，合計で 549 枚の舌画像を利用した．なお、このデータは千葉大学大学院医学研究院倫理審査承認番号 812 号の承認を得ている．これらの舌画像とペアで学習させる二値化舌画像の作成にはフリーソフトを利用し，舌の領域を手作業で

なぞり，その領域内を白く，それ以外を黒くする処理を行うことで作成した．図 3 は今回作成した学習データのペアの一例である．

学習させる際の各種条件は表 1 の通りである．データ数 549 ペアの内 9 割を訓練データ，1 割をテストデータとし 4000 エポック学習させる．学習させる画像サイズは 256×256pixel である．バッチ数は 4 ペアとし，5000 イテレータごとにパラメータを保存するが，保存した全てのパラメータにおいて画像を生成する．生成された画像はグレースケールであるため 0 以上 200 未満を 0 に，200 以上を 255 に変換する処理を行う．表 2 は実験を行った環境である．

表 1 各種条件

データ数（pair）	549
エポック数（epoch）	4000
バッチサイズ（batch）	4
パラメータ保存間隔(iterator)	5000

表 2 実験環境

OS	CentOS
メモリ	60GB
プロセッサ	Intel(R)core(TM)i7-5820K CPU @ 3.30GHz
GPU プロセッサ	NVIDIA GeForce GTX TITAN X 12GB

4. 結果

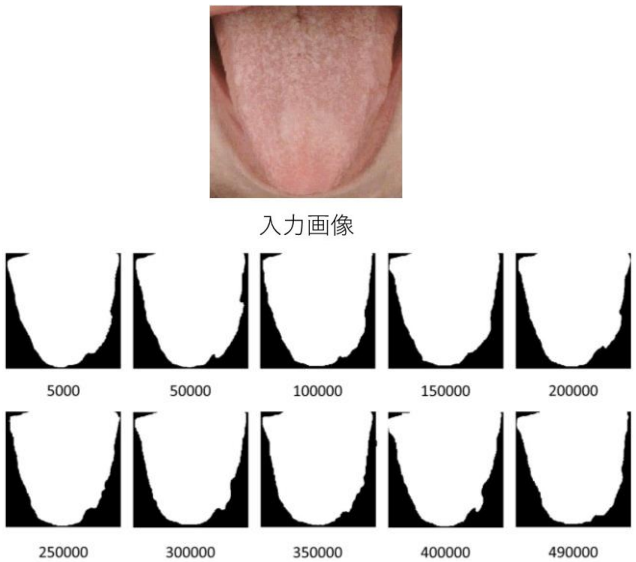

図 4 イテレータごとの生成画像

保存した各パラメータおいて画像を生成させたところ図 3 のような二値化舌画像が生成された．学習が進んでも，白く指定すべき舌の領域が安定せず，最後まで収束しなかった．これは訓練データとなる二値化画像を作成した際に手作業で行ったために，舌とその他の部分との境界があいまいになったことが考えられる．

ここで，どの生成画像においても実際の舌の領域は白く指定されていることから，全ての生成画像の白く指定した領域の中で，一番内側のものを舌の領域として採用することで，一枚の二値化舌画像を作成することにした．この画像は，図 4 のようにピクセルごとの論理積を計算することで作成した．

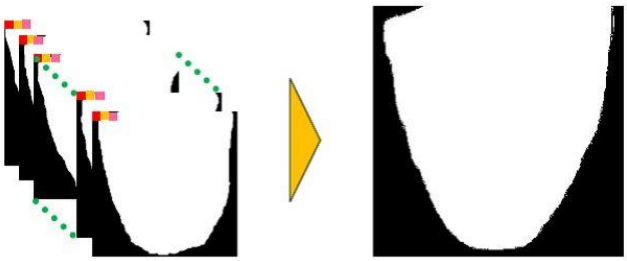

図 5 ピクセルごとの論理積を計算

49 人の舌画像から二値化舌画像を生成した．学習に要した時間は約 3 日，1 枚の二値化画像の生成に要した時間は約 10 分であった．二値化画像の生成に時間がかかったのは，現状では保存した全てのパラメータにおいて画像を生成し，生成した全ての画像の論理積を取って 1 枚の画像としているためだと考えられる．

入力した舌画像と生成された二値化舌画像，生成した二値化舌画像と入力した舌画像の舌領域の差異を図 6 に示す．

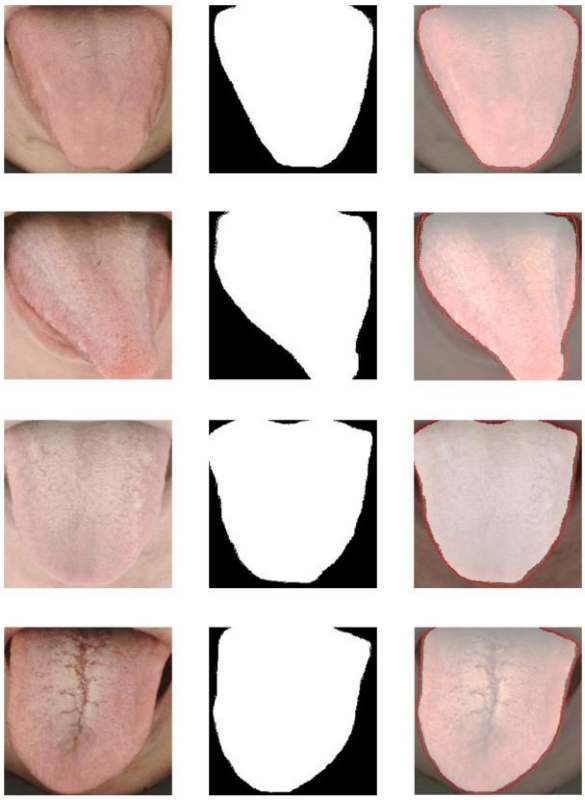

図 6　生成した二値化舌画像
左から入力画像，生成した二値化舌画像，入力画像の舌領域との差異

生成した二値化舌画像の舌領域は，入力した舌画像の舌領域よりも若干ではあるが小さくなった．これは論理積を取ったことにより，より内側のものが採用されたことが原因だと考えられる．しかし，手作業で作成した二値化舌画像と遜色のないものが作成できたように見られる．そこで，手作業で作成した二値化舌画像と生成した二値化舌画像の類似度を Dice 係数を利用して評価する．Dice 係数は以下の式で定義される．

$$\mathrm{DSC}(X,Y) = \frac{2|X \cap Y|}{|X| + |Y|},$$

これは２つの集合 X と Y の平均要素数と共通要素数の割合を表している．DICE 係数の値が 1 に近いほど，２つの集合の類似度は高いと言える．この DICE 係数を利用して類似度を算出した結果，最高値 0.98 と最低値 0.93，平均値 0.96 であった．この結果から精度のよい二値化舌画像が生成できたと考えられる．

多くの舌画像において精度の良い二値化舌画像が生成できたが，一部には図 7 のように舌領域を白く指定できていないものや，舌以外の場所を白く指定しているものがあった．これらは影になって舌が暗くなっている部分に多く，暗い部分が舌であるのかどうかという学習データの不足が一つの原因として考えられる．

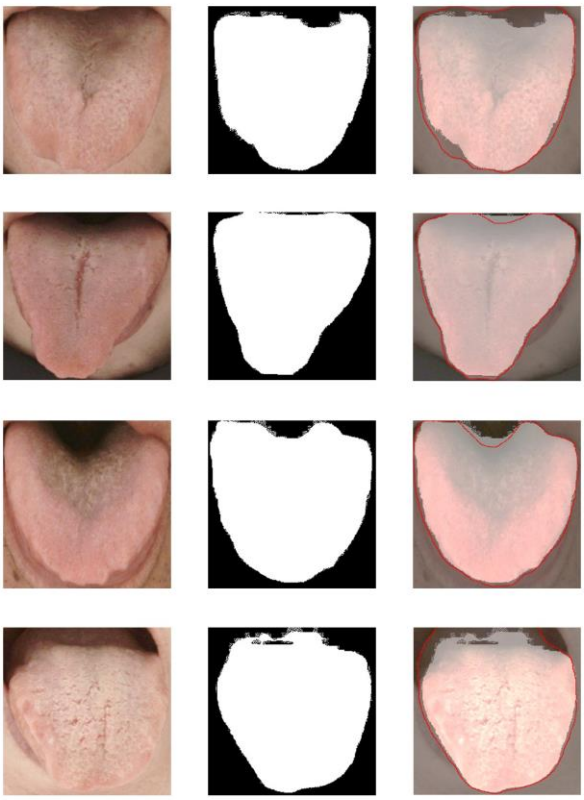

図 7 二値化精度の悪い二値化舌画像
左から入力画像，生成した二値化舌画像，入力画像の舌領域との差異

5. まとめと今後の展望

5.1 まとめ

本研究では，標準舌形状に統一する手法の一つである TPS を行う際に，重心を求めるために必要となる二値化舌画像を自動生成させる手段として GAN の一種である pix2pix を利用した．生成した二値化舌画像の舌領域は，実際の舌領域よりも若干小さくなったが，手作業で作成したものと遜色のない精度で生成できた．しかし，生成した二値化舌画像の一部には，舌領域を白く指定しきれていないものや，舌以外の場所を白く指定しているものがあった．

5.2 今後の展望

549 ペアのデータセットを利用したが，実質は 39 人分という少ない舌画像の種類を学習させるだけで高い精度の二値化舌画像を生成できたため，より学習データを増やすことで精度を向上させられると考えている．また，今回使用した画像は同一条件下で撮影されたものであるため，違う条件で撮影された舌画像に対しては，精度の良い二値化舌画像を生成できない可能性があるため，様々な条件下で撮影した舌画像を学習させる必要もある．その他，二値化舌画像の生成に約 10 分と長い時間がかかっていることから，論理積に用いる画像枚数の検証を行うことや，より高い精度で二値化を行うための教師データと見直しなどが挙げられる．また今回利用した pix2pix は多種多様な画像生成に利用できる汎用的なアルゴリズムであるため，セグメンテーションに特化したアルゴリズムを用いた際の精度の違いも確かめたい．

参考文献

[1] Chuang-ChienChiu, “A novel approach based on computerized image analysis for traditional Chinese medical diagnosis of the tongue”, Computer Methods and Programs in Biomedicine, vol. 61, pp77-89 (2000).
[2] F.L.Bookstein, “Principal warps: Thin-plate splines and the decomposition of deformations“, IEEE Trans. Pattern Anal. Mach Intel, vol.11, pp567-585 (1989).
[3] Mohammad Havaei, Axel Davy, David Warde-Farley, Antoine Biard, Aaron Courville, Yoshua Bengio, Chris Pal, Pierre-Marc Jodoin, Hugo Larochelle “Brain tumor segmentation with Deep Neural Networks”, ArXiv, [Online]. Available: arXiv:1505.03540, to be published (2015).
[4] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, Alexei A. Efros, “Image-to-Image Translation with Conditional Adversarial Networks”, In CVPR (2017). https://phillipi.github.io/pix2pix/
[5] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, Yoshua Bengio, ” Generative Adversarial Nets”, Advances in Neural Information Processing Systems 27 (2014)
[6] Olaf Ronneberger, Philipp Fischer, Thomas Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation"In Miccai, Volume 9351, pp234-241(2015)