

弱教師あり学習による連続的な表情特徴の獲得

狩野 悌久^{1,a)} 長尾 智晴^{2,b)}

受付日 2021年9月13日, 再受付日 2021年11月6日,
採録日 2021年11月26日

概要: 機械学習による表情特徴の獲得は, 一般的に感情ラベルなどを用いた教師あり学習によって行われる。しかし, 人間による表情のラベリングは, 主観的なものになりやすく教師自体が曖昧性を持つ可能性がある。また, 表情に対して感情のクラスを割り当てることは連続的な表情を離散的に扱うことになり, モデルが表情の連続性を学習することを妨げる。そこで私たちは, 表情に関連する教師情報を用いることなく, 被験者特徴から切り離された状態の表情特徴を獲得することを研究の目的とした。本稿では, 以前私たちが提案した手法に対して2種類の損失関数を導入し, さらに学習プロセスを改良することにより, しわなどの細かな情報を含んだ表情特徴を獲得する手法を提案する。実験では, 表情認識と画像生成を行い, 提案手法によって獲得された特徴量の有効性を示す。

キーワード: 弱教師あり学習, 繚れを解いた特徴表現学習, 表情認識, 画像生成, 変分オートエンコーダ, 敵対的生成ネットワーク

Weakly Supervised Learning for Acquisition of Continuous Facial Expression Features

YOSHIHISA KANOU^{1,a)} TOMOHARU NAGAO^{2,b)}

Received: September 13, 2021, Revised: November 6, 2021,
Accepted: November 26, 2021

Abstract: Acquisition of facial expression features from images by machine learning is generally done with supervision, such as using emotional labels that match facial expressions. However, in the supervised setting, there are problems such as ambiguity of the label due to the subjectivity of the person to the facial expression and the discrete treatment of continuous facial expressions by giving the label. To solve these problems, we focus on acquiring continuous facial expression features without using information of facial expression for training the model. In this paper, we improve the weakly supervised method proposed in “Separation of the Latent Representations into Identity and Expression without Emotional Labels” to acquire more effective facial expression features. The experimental result shows that the proposed method acquire effective facial expression features and achieve better results than the previous method in each task of image generation and facial expression recognition.

Keywords: weakly supervised learning, disentangled representation learning, facial expression recognition, image generation, variational autoencoder, generative adversarial network

¹ 横浜国立大学大学院環境情報学府
Graduate School of Environment and Information Sciences,
Yokohama National University, Yokohama, Kanagawa 240–
8501, Japan

² 横浜国立大学大学院環境情報研究院
Faculty of Environment and Information Sciences, Yoko-
hama National University, Yokohama, Kanagawa 240–8501,
Japan

^{a)} kanou-yoshihisa-wm@ynu.jp

^{b)} nagao@ynu.ac.jp

1. はじめに

機械学習, 特に深層学習において, タスクに対して効果的な特徴を獲得し, モデルのロバスト性を高めるためには, 大量の教師ありデータを学習に利用する必要がある [1], [2]. これは, 表情特徴の獲得についても同様であり, 画像からの表情認識 [3], [4] や顔画像生成 [5] を行う研究では, Ekman の基本六感情 [6] やこれを拡張した感情 [7], [8]

など、表情に紐づいた情報をクラスとして学習に利用することが多い。しかし、人による表情の認識は主観的なものであるため、教師として与えたクラス自体が曖昧性を含む可能性がある。また、表情に対してラベリングを行い、いくつかのクラスとして扱うことは、連続的な表情を離散的に扱うことになり、モデルが表情の連続性を維持した特徴を獲得することを妨げてしまう。

これらの問題を解決するために、先行研究 [9] では、感情ラベルのような表情に関連する情報を利用せず、付加情報としては被験者情報（被験者 ID）のみを学習に利用することで、Variational autoencoder (VAE) [10], [11] の潜在変数として、表情特徴と被験者特徴を分離した状態で獲得する手法を提案した。しかし、この手法では、VAE の特性から生成画像がぼやけたものになり、しわなどの細かな部分の再構築は実現できていなかった。またこの点から、潜在変数として獲得された表情特徴についても、詳細な表情は十分に表現できていないことが示唆されていた。

そこで本稿では、この手法を改良し、被験者情報を用いた弱教師あり学習によって、より効果的な表情特徴を獲得する手法を提案する。提案手法では、2 種類の損失関数の導入と学習ステップの改善を行うことで、被験者特徴と表情特徴を分離しつつ、詳細な表情特徴の獲得を行う。実験では、モデル潜在空間上でのユークリッド距離を用いた表情認識と 2 つの顔画像生成タスクを実施し、従来手法と比較を行うことで、提案手法の有効性を検証する。

2. 関連研究

機械学習で特徴獲得を行う際、様々な要素が複雑に絡みあうことなく、人間にとって意味的に解釈可能な特徴として獲得することは、Disentangled representation learning (纏れを解いた表現学習) と呼ばれ、様々なドメインにおいて研究がなされている。コンピュータビジョンの分野においてこのような表現の学習は、画像認識のタスクに対して有効な特徴を認識に用いることを目的としたり、画像生成のタスクに対して生成画像に現れる属性を制御するために利用される。また、纏れを解いた表現学習は、教師なしで行われる場合とラベルなど付加情報を利用して教師ありや弱教師ありの枠組みで行われる場合が存在する。

infoGAN [12] や β -VAE [13] は教師なし学習で、解釈可能な表現を学習する手法として知られている。infoGAN は、入力ノイズを解釈可能な意味を持たせる潜在変数 C とそれ以外の要素を構築するためのノイズ z に分解し、 C と生成された画像の間の相互情報量を最大化するように学習を行うことにより、潜在変数 C に解釈可能な表現を獲得する手法である。 β -VAE は、VAE の近似事後分布を事前分布に近づける正則化項 β を導入したものであり、潜在空間に対して制限を加えることで、纏れを解いた表現を獲得する手法である。しかし、これらの手法は、Mnisit [14] な

どの比較的単純なデータの場合では良好な結果を示すが、CelebA [15] など複雑なデータに適用した際に、特徴の分離がうまく行えない問題が存在する。

そこで、Liu らは被験者情報を用いることで、顔画像から眼鏡の有無や男性女性などの属性特徴と被験者特徴を分けて獲得する D^2 -AE [16] を提案した。また、Bouchacourt らは、データを複数のグループとして見た際に、グループ内で共有する特徴 (style) と共通しない特徴 (content) に分けて、モデルの潜在表現として獲得する ML-VAE [17] を提案している。これらの手法が完全な教師なし学習ではなく、付加情報を利用した弱教師あり学習であるため、複雑なデータであっても比較的良好に画像生成を行うことが可能である。しかし、これらの方法によって獲得された特徴は複数の属性が混ざりあった状態で獲得されるため、実際に潜在変数と画像に現れる各属性の対応付けを行う作業は人間の目視によって行う必要がある。また、従来の研究の多くは眼鏡の有無などの離散的な属性値を主に対象としており、表情などの連続的なイベントは、笑顔など部分的には扱われているがその連続性は考慮されていない。

3. 先行研究: Separation of the Latent Representations into Identity and Expression without Emotional Labels

先行研究 [9] は VAE を拡張し、被験者 ID を利用した 2 段階の学習により、潜在変数に表情特徴と被験者特徴を分離した状態で獲得する手法である。モデル構造は、図 1 (a) に示すように 2 つのエンコーダと共通する 1 つのデコーダによって構成されており、それぞれのエンコーダは被験者特徴と表情特徴をそれぞれの潜在変数 z_i , z_e に埋め込むことを目的としている。エンコーダにこれらの機能を具備するために、モデルの学習は被験者特徴の獲得を目的としたステージと表情特徴の獲得を目的としたステージの 2 段階で行われる。

被験者特徴の獲得を行うステージでは、被験者 ID を利用し、IdentityEncoder ($q(z_i|x; \phi_i)$) と Decoder ($p(x|z; \theta)$) の学習が行われる。通常の VAE の学習では、入力画像 x とモデル出力 x' の間で再構築誤差が計算されるのに対し、このステージでは、図 1 (b) および式 (1) の第 2 項に示すように、入力した画像と同じ被験者の画像集合 X_{id} の中からランダムにサンプリングされた画像 $x_{rand_{id}}$ とモデル出力の間で再構築誤差の計算を行う。

$$\mathcal{L}_i = \alpha_{i1} D_{KL}(q(z_i|x; \phi_i)||p(z)) - \alpha_{i2} E_{q(z|x; \phi_i)}[\log p_{\theta}(x_{rand_{id}}|z)] \quad (1)$$

式 (1) の α_{i1} , α_{i2} はそれぞれの項に対する重みパラメータであり、 z は z_i , z_e を結合したベクトルを表している。また、このステージでは ExpressionEncoder の学習は行われないため、 z_e として潜在変数と同じ次元数のゼロベクトル

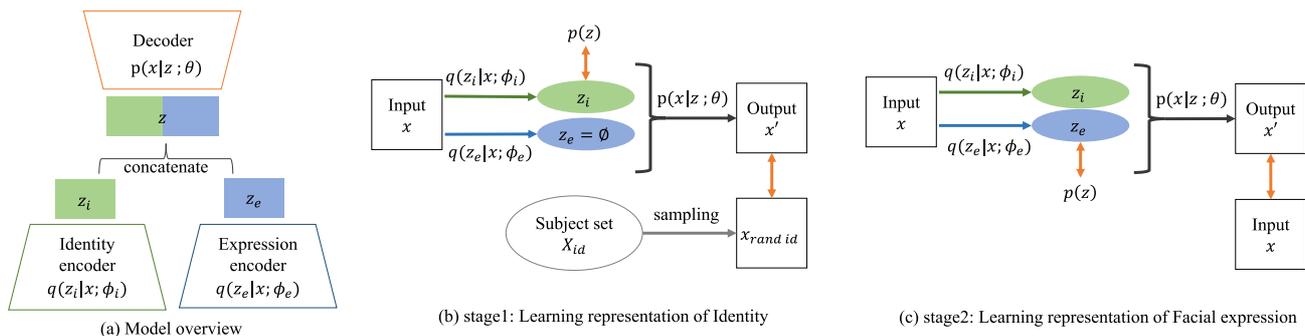


図 1 従来手法：モデル構造および学習ステージ

Fig. 1 (a) represents overview of the model and (b), (c) shows training steps. (b) shows the learning identity representation stage. In this stage, the ExpressionEncoder is not update, and reconstruction error is calculated between randomly sampled image from the same subject's image set $x_{rand_{id}}$ (id is subject's identity number) and output. (c) shows the learning facial expression representation stage. In this stage the IdentityEncoder is not update, and reconstruction error is calculated between input and output same as VAE.

ルが代入される．これは，VAE では潜在変数空間の事前分布としてガウス分布 $\mathcal{N}(0, I)$ を仮定しており，潜在変数において最も平均的な状態は 0 であることから，ゼロベクトルを代入することで被験者特徴獲得時の表情特徴を最も平均的な状態として扱うためである．この損失関数は，同じ被験者集合からランダムにサンプリングされた画像を使用して再構成誤差を計算するため，入力画像の表情を考慮しない学習が進められる．また，モデルは同じ被験者のすべてのデータの組合せの間で再構成誤差を減少させる必要があるため，各被験者の最も尤度の高い画像を出力するように学習が行われる．このとき，IdentityEncoder の潜在変数 z_i には，入力画像の表情に影響されることのない特定の被験者に対して最も尤度の高い状態，つまり被験者特徴が埋め込まれることが期待される．このようにして，IdentityEncoder は機能獲得を行う．

次に，表情特徴の獲得を行うステージでは，IdentityEncoder のパラメータ ϕ_i は固定し，ExpressionEncoder ($q(z_e|x; \phi_e)$) と Decoder ($p(x|z; \theta)$) の学習が行われる．このステージでは，前段の被験者特徴獲得のステージとは異なり，通常の VAE と同様の損失関数 (式 (2)) により，入力画像 x とモデル出力 x' の間で再構築誤差を計算する．

$$\mathcal{L}_e = \alpha_{e1} D_{KL}(q(z_e|x; \phi_e) || p(z)) - \alpha_{e2} E_{q(z|x; \phi_e)} [\log p_\theta(x|z)] \quad (2)$$

前段の学習により，すでに被験者特徴は潜在変数 z_i に埋め込まれているため，モデルは ExpressionEncoder の潜在変数 z_e に対して， z_i に不足する要素を埋め込むことにより，入力画像を再構築するように学習される．ここで， z_i に不足する要素は表情特徴であることが期待されるため，ExpressionEncoder は表情特徴を抽出する機能を獲得することになる．また，この手法では VAE のように

Reparameterization Trick を利用して，学習時の潜在変数を決定しているため，潜在空間上で近い点に位置する表現は似通った表現になりやすくなる．その結果，ExpressionEncoder の潜在空間では，被験者が異なる場合であっても類似する表情は近い点にプロットされる．このようにして，ExpressionEncoder は連続的な表情特徴を獲得する．

先行研究では，このように学習の制御に被験者 ID を利用した 2 段階の学習を行うことで，連続的な表情特徴の獲得を実現している．

4. 提案手法

先行研究では VAE の特性上，生成画像がぼやけてしまうため，顔のしわなどの表情を構築するうえで重要な，細かな特徴がとらえられていないことが示唆されていた．生成画像を鮮明化するアプローチとしては，再構築誤差の影響を大きくする方法やモデルのパラメータ数を増やす方法などが存在する．しかし，それらの方法では過学習が生じることでテストデータに対する再構築結果が良好でなくなる場合や，被験者特徴と表情特徴が絡み合った特徴として獲得される可能性がある．そこで提案手法では，2 種類の損失関数を新たに導入し，さらに学習ステップを改良することにより，モデルの過学習を抑えつつ詳細な顔画像生成を行うことで，より効果的な表情特徴の獲得を行う．

4.1 損失関数の改良

ここでは，より詳細でかつ被験者に共通する表情特徴の獲得を目的として導入された，2 つの損失関数について説明する．これらの関数は表情特徴の獲得に対して効果を期待する損失関数であるため，被験者特徴獲得のステージには利用せず，表情特徴の獲得のステージにのみ導入する．

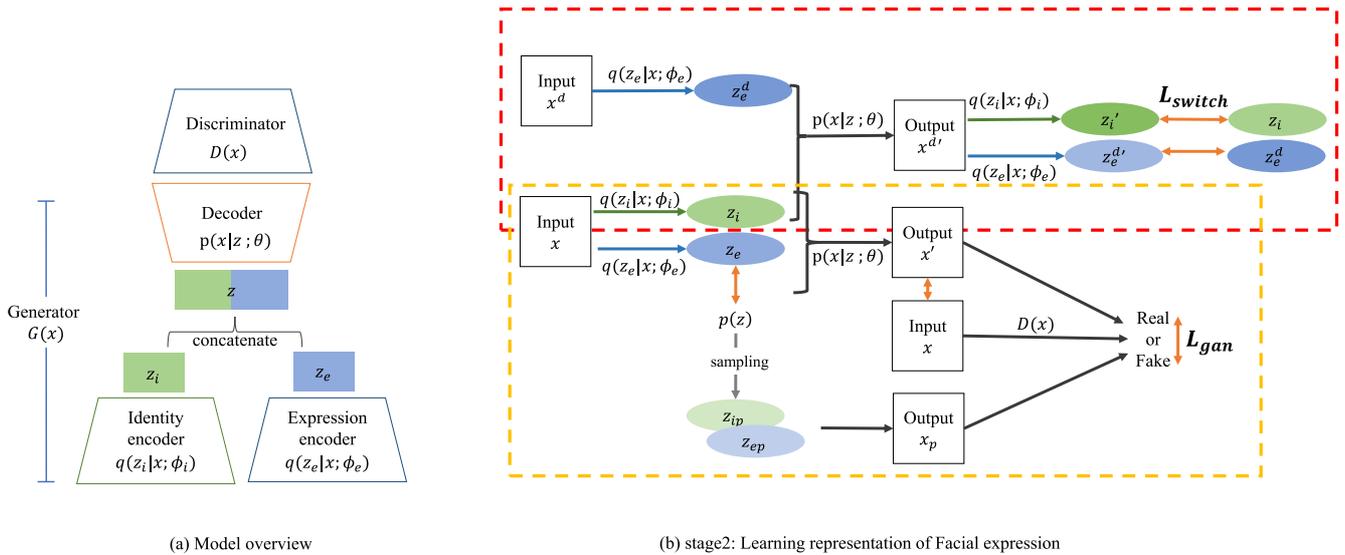


図 2 提案手法：モデル構造および表情特徴獲得ステージ

Fig. 2 (a) represents overview of our model and (b) shows learning step of acquiring facial expression representation. In the proposed method, Discriminator is introduced into the model and two types of loss functions, L_{gan} and L_{switch} , are added.

4.1.1 Adversarial Loss

図 2 (b) の黄色の点線で囲われた部分は、今回導入した Adversarial loss の概要を表している。Adversarial loss は GAN [18] にて提案された損失関数であり、Generator と Discriminator の 2 種類のネットワークを利用して生成を行う際に用いられる。画像生成のタスクにおいては、Generator は画像生成を行うネットワークであり、Discriminator は入力された画像が、Generator によって生成された画像か本物の画像かを判断するネットワークである。これらのネットワークが互いに影響を与えながら学習が行われることより、Generator は Discriminator を騙すように学習が行われるため、結果として Generator は本物に近い、鮮明な画像を生成することができる。この性質を利用し、生成する顔画像を鮮明化することで、より詳細な表情の特徴を潜在変数に獲得する。提案手法では、図 2(a) に示すように 2 つの Encoder と Decoder を Generator (G) として扱い、新たに Discriminator (D) を追加する。

$$\mathcal{L}_{gan} = \min_G \max_D \{ \log(D(x)) + \log(1 - D(x')) + \log(1 - D(x_p)) \} \quad (3)$$

式 (3) は提案手法で導入される Adversarial loss を表したものであり、 x はデータサンプル（本物の画像）を表し、 x' は Generator の出力を表している。また、 x_p は、VAE の事前分布 $\mathcal{N}(0, I)$ からサンプリングされた値を z として Decoder に入力した際の出力である。

4.1.2 Switch Loss

図 2 (b) の赤色の点線で囲われた部分は Switch loss の概要を表したものである。Switch loss は被験者特徴と表情特

徴が絡み合った特徴として獲得されることを抑制し、被験者間で共通する表情特徴の獲得を目的として導入する。ここで、被験者間で共通する表情特徴とは、被験者が異なる場合であっても、似通った表情は潜在変数として近い値をとる特徴量のことである。Adversarial loss のみを表情特徴獲得のステージに導入した場合では、モデルは画像の再構築だけではなく鮮明化を目的として学習を行う。そのため、被験者特徴獲得のステージで獲得された特徴を無視した学習が行われやすくなり、表情特徴が被験者固有のもの、もしくは被験者特徴と絡み合った特徴になることが予測される。そこで、被験者に共通した表情特徴を獲得するために、式 (4) に示す損失関数を導入する。

$$\mathcal{L}_{switch} = MSE(z_i, z_i') + MSE(z_e^d, z_e^d) \quad (4)$$

ここで z_i はある被験者の画像から IdentityEncoder によって抽出された被験者特徴を表し、 z_e^d は z_i とは異なる被験者の画像から ExpressionEncoder によって抽出された表情特徴を表している。また、 z_i' 、 z_e^d はそれぞれ、 $z = (z_i, z_e^d)$ を Decoder に入力した際に得られた出力を再度 IdentityEncoder, ExpressionEncoder に入力した際の潜在変数を表している。つまり、この損失関数は、異なる被験者間で表情特徴が入れ替えられた場合においても、Decoder の出力にその表情特徴を維持するように設計された関数である。また、被験者特徴である z_i と z_i' の間でも誤差を取っていることから、モデルが被験者特徴獲得のステージで獲得した特徴を無視して学習を行うことを抑制する。これにより、獲得される表情特徴が被験者間で共通の特徴となることが期待できる。

表情特徴獲得のステージでは 4.1.1, 4.1.2 項で説明した損失関数と式 (2) を組み合わせ、最終的に下記目的関数 \mathcal{L}'_e に従って学習を行う。

$$\mathcal{L}'_e = \mathcal{L}_e + \beta \mathcal{L}_{gan} + \gamma \mathcal{L}_{switch}, \quad \beta, \gamma > 0 \quad (5)$$

4.1.3 学習ステップの改良

ExpressionEncoder の潜在変数 z_e と IdentityEncoder の潜在変数 z_i は互いに異なる情報を保持し、情報を補完し合うことで顔画像の生成を可能にしている。そのため、それぞれのエンコーダの学習は、互いに影響を与えながら行われることが望ましいといえる。しかし、従来手法の学習ステップでは、それぞれのステージで一方のエンコーダのみパラメータの更新が行われ、また学習は一巡のみしか行われないうちに、第 1 ステージで学習される IdentityEncoder は ExpressionEncoder の学習結果を加味することができていなかった。そこで我々は、被験者特徴獲得のステージと表情特徴獲得のステージを繰り返すことで、IdentityEncoder の学習時に、ExpressionEncoder の影響を間接的に与えることを行う。ここで“間接的”という言葉を用いる理由は、被験者特徴獲得ステージでは ExpressionEncoder の潜在変数は利用されず、 z_e にはゼロベクトルが代入されるため、ExpressionEncoder の学習結果は Decoder のパラメータによって、IdentityEncoder に伝えられるためである。

また、このステージの繰り返し処理は、Adversarial loss と組み合わせることにより被験者特徴と表情特徴を分離した状態で、生成画像を鮮明化する効果も期待できる。一般に、Adversarial loss を利用して鮮明な画像を生成しようとする際には、学習エポック数を増やす必要がある。しかし、一巡の学習ステップでそれを行おうとすると、表情特徴の獲得ステージのエポック数を増やす必要があるため、ExpressionEncoder と Decoder が過学習を起し、 z_e のみに依存して Decoder が画像生成を行ってしまう可能性がある。一方、ステージを繰り返して学習を行う場合は、表情特徴の獲得ステージのエポック数を増やしつつ、IdentityEncoder の学習も逐次行われるため、過学習が抑制され特徴が纏れを解いた状態で獲得することが可能になる。

5. 実験

この章では、提案手法により効果的な表情特徴の獲得が実現できることを検証するために、表情認識と画像生成の 2 種類タスクを対象として実験を行う。表情認識のタスクでは、ExpressionEncoder の潜在空間を利用してユークリッド距離によるクラスタリングを行うことで、表情特徴が被験者間で共通のものとして獲得されているか評価を行う。また、画像生成のタスクでは、被験者間で表情の入れ替えを行うことで被験者特徴と表情特徴の分離度を評価し、中

間表情の生成を行うことで獲得された表情特徴の連続性を評価する。

5.1 データセットおよび実験設定

この実験では下記の 3 つのデータセットを組み合わせ利用する。

MUG

MUG Facial Expression Database [19] は 20 歳から 35 歳の 86 人の被験者 (男性 51 名, 女性 35 名) で構成されたデータセットである。それぞれのデータは、Neutral な表情から、Ekman の基本六感情 (anger, disgust, fear, happiness, sadness, surprise) の 1 つの感情を表情で表現し、再び Neutral に戻る画像列である。

CK+

The Cohn-Kanade database [20] は、18 歳から 50 歳の 123 人の被験者、593 個の画像列から構成されたデータセットである。データは実験環境にて撮影されたものであり、それぞれのデータには MUG と同様に Ekman の基本六感情が割り当てられている。

RAVDESS

The Ryerson Audio-Visual Database of Emotional Speech and Song [21] は、21 歳から 33 歳までの 24 人の被験者 (男性 12 人と女性 12 人) の画像列と音声で構成されたデータセットである。それぞれのデータは、俳優が基本六感情に Neutral, calmness を加えた 8 つの感情を、演技によって表現しながら歌った様子を記録した動画である。

実験ではこのデータセットからそれぞれ 2 人～3 人の被験者テスト用の被験者とし、残りを学習用の被験者とした。また、実験用のデータセットとして、モデル学習用と表情認識評価用の 2 種類のデータセットを構築する。モデル学習用のデータセットは、学習用の被験者の各画像列から等間隔で画像のサンプリングを行い作成する。これにより、表情が変化する途中の微妙な表情をデータセットに含めることが可能になり、表情の多様性を確保される。表情認識評価用のデータセットは、画像の表情と感情ラベルが一致する必要があるため、学習用、テスト用両方の被験者の画像列からデータに割り当てられた表情が出現する部分を自動的に抽出し作成する。つまり、MUG では中心の前後の数フレームを CK+ では最後の数フレームを利用し、RAVDESS は感情に紐づいた表情が表出する箇所を自動で決定できないため利用しないこととした。

最終的に、学習用の被験者は 195 名、テスト用の被験者は 7 名であり、モデル学習用のデータセットは 29,085 枚の画像、表情認識評価用のデータセットは 2,037 枚 (学習用被験者: 1,842 枚, テスト用被験者: 195 枚) の画像を有するデータセットとして構成された。また画像は、学習済みの顔検出モデル [22] を使用して顔領域をトリミングし、

表 1 エンコーダの構造

Table 1 Structure details of encoder.

Type	Ksize	Stride	Pad	Output
Image data	-	-	-	$3 \times 64 \times 64$ ($3 \times 128 \times 128$)
conv1.1	3×3	2	1	$32 \times 32 \times 32$ ($32 \times 64 \times 64$)
conv1.2	3×3	1	1	$32 \times 32 \times 32$ ($32 \times 64 \times 64$)
conv2.1	3×3	2	1	$64 \times 16 \times 16$ ($64 \times 32 \times 32$)
conv2.2	3×3	1	1	$64 \times 16 \times 16$ ($64 \times 32 \times 32$)
conv3.1	3×3	2	1	$128 \times 8 \times 8$ ($128 \times 16 \times 16$)
conv3.2	3×3	1	1	$128 \times 8 \times 8$ ($128 \times 16 \times 16$)
conv4.1	3×3	2	1	$256 \times 4 \times 4$ ($256 \times 8 \times 8$)
conv4.2	3×3	1	1	$256 \times 4 \times 4$ ($256 \times 8 \times 8$)
(conv5.1)	3×3	2	1	($256 \times 4 \times 4$)
(conv5.2)	3×3	1	1	($256 \times 4 \times 4$)
average pooling	4×4	1	1	$256 \times 1 \times 1$
fc. μ	-	-	-	64
fc. σ	-	-	-	64

表 2 デコーダの構造

Table 2 Structure details of decoder.

Type	Ksize	Stride	Pad	Output
latent variable	-	-	-	128
fc1	-	-	-	4096
reshape	-	-	-	$256 \times 4 \times 4$
deconv1.1	4×4	2	1	$128 \times 8 \times 8$
conv1.2	3×3	1	1	$128 \times 8 \times 8$
deconv2.1	4×4	2	1	$64 \times 16 \times 16$
conv2.2	3×3	1	1	$64 \times 16 \times 16$
deconv3.1	3×3	2	1	$32 \times 32 \times 32$
conv3.2	3×3	1	1	$32 \times 32 \times 32$
deconv4.1	3×3	2	1	$16 \times 64 \times 64$
conv4.2	3×3	1	1	$16 \times 64 \times 64$
(deconv5.1)	3×3	2	1	($8 \times 128 \times 128$)
(conv5.2)	3×3	1	1	($8 \times 128 \times 128$)
conv6	3×3	1	1	$3 \times 64 \times 64$ ($3 \times 128 \times 128$)

$64 \times 64 \times 3$ または $128 \times 128 \times 3$ にリサイズされた後、モデルに入力される。

エンコーダとデコーダの構造をそれぞれ表 1, 表 2 に示す。括弧で示した箇所は、入力サイズが $3 \times 128 \times 128$ のときに追加で用いるレイヤー、出力サイズを表している。また、ExpressionEncoder と IdentityEncoder の潜在変数の大きさはともに 64 次元に設定した。Discriminator の構造は出力層までを Encoder と同様とし、出力層の出力サイズのみ 1 次元に変更した。モデルのパラメータは He が提案した手法 [23] を用いて初期化され、最適化手法には Adam [24] ($\beta_1 = 0.9$, $\beta_2 = 0.999$, $\sigma = 1.0 \times 10^{-8}$, $lr = 0.0005$) を用いた。学習エポック数はそれぞれのステージで 100 epoch (繰り返し学習を行う場合は 50 epoch) とし、学習率はエポックごとに 0.95 を乗算し徐々に減衰させた。また、目的関数 L'_e のパラメータは実験的に $\alpha_{i1} = 1.0 \times 10^5$, $\alpha_{e1} = 1.0 \times 10^5$, $\alpha_{i2} = 0.01$, $\alpha_{e2} = 0.01$, $\beta = 1$, $\gamma = 10$ とした。

5.2 実験結果および考察

5.2.1 表情認識

この実験では、ExpressionEncoder の潜在空間を利用

し、ユークリッド距離によるクラスタリング手法 (k-means++ [25]) によって表情認識を行うことで、獲得された表情特徴の評価を行う。提案手法により、ExpressionEncoder が被験者に共通した表情特徴の獲得が行えている場合、似通った表情の潜在空間上でのユークリッド距離は、仮に被験者が異なっている場合であっても小さくなるため、表情認識の結果が良好になることが期待される。今回の実験では、kmeans++ のクラスタ数を $k = 9$ とし、クラスタへのクラスの決定にはセントロイドから最も近いポイントの画像のラベルを利用した。また k-means++ は初期のセントロイドの決定がランダムに行われ、それにより精度が変わるため、クラスタリングは 100 施行し平均を取った。

比較手法としては、まずベースラインとなる精度を確認するために、オリジナルな VAE の潜在空間を利用した場合を実施した。また、被験者情報を利用した場合の既存手法による表情認識の精度を確認するため、被験者 ID をラベル情報として利用した場合の C-VAE [26] の潜在空間を用いた場合と、オリジナルな VAE の潜在変数に対して式 (6) で示すように被験者ごとに潜在変数の平均をとり、それぞれから減算することで、被験者特徴を潜在変数から除く処理を加えた場合の実験を実施した。

$$f_{sub_{ij}} = x_{ij} - \frac{1}{n} \sum_{k=1}^n x_{ik} \quad (6)$$

i : subject identity number

x_{ij} : the j th image in the subject set X_i

表 3 はそれぞれの表情認識の結果を表している。ここで、train はモデル学習に利用した被験者のデータ、test は学習に利用していない被験者のデータを表しているため、分類結果は train, test とともに教師なしで行われた結果である。また、⑦⑧に test の結果がない理由は、test において被験者 ID は未知であることを前提としているが、これらの手法では被験者 ID をモデルの入力もしくは特徴量の算出に利用しているため、テストデータに対して特徴量の算出が行えないためである。

提案手法④と従来手法①、および比較手法⑥⑦⑧の結果を比較すると、提案手法が最も高精度に表情の認識を行えていることが分かる。また、②と③の結果から、Adversarial loss のみを導入した場合には、IdentityEncoder を無視し ExpressionEncoder と Decoder のみで画像の再構築が行われたため、被験者特徴と表情特徴の分離がうまく行えず、精度が従来手法よりも低い結果になっていたが、Switch loss を導入することでその問題が解決され、精度向上に転じていることがうかがえる。さらに、提案手法では Adversarial loss を導入し、画像の鮮明化による細かな表情特徴の獲得を行っているため、入力サイズを 64 から 128 に大きくすることで、より細かな表情情報を学習に利用することが可

表 3 表情認識結果

Table 3 Results of facial expression recognition.

	input size	train	test
①conventional method [9]	64	56.33 ± 2.54%	49.95 ± 3.58%
②add L_{gan}	64	53.46 ± 2.66%	48.10 ± 3.26%
③add $L_{gan}&L_{switch}$	64	62.64 ± 2.48%	59.95 ± 4.31%
④add $L_{gan}&L_{switch}$ (iterate Stage)	64	64.28 ± 2.81%	62.83 ± 5.24%
⑤add $L_{gan}&L_{switch}$ (iterate Stage)	128	67.95 ± 3.71%	65.92 ± 6.65%
⑥original VAE	64	38.08 ± 1.56%	22.07 ± 3.21%
⑦original VAE (subtract subject mean)	64	48.99 ± 2.76%	-
⑧C-VAE	64	59.05 ± 4.23%	-

能となり、表情認識に対して有効な特徴を獲得できていることが、⑤の結果で示されている。

学習ステージ繰り返しの影響に着目すると、④の結果が③を上回っていることから、学習ステージの繰り返しにより、Encoder が互いに影響を与えながら学習が進み、より効果的な表情特徴の獲得を実現しているといえる。また、図 3 は入力サイズが 128 の場合における、学習ステージの繰り返しによる認識精度の変化を表したグラフである。ここから、繰り返し学習を行うことで徐々に精度が向上していき、1 巡で学習した場合よりも高い精度を実現していることが分かる。これは、それぞれの学習ステージを交互に行うことにより、それぞれの Encoder が互いに影響を与えながら学習することが可能となり、より明確に被験者特徴と表情特徴を分離するように学習が進んだためであると考察される。

以上の結果から、提案手法の行った 2 種類の損失関数の導入と学習ステージの繰り返しにより、被験者に依存しないより効果的な表情特徴の獲得が行えていることが確認された。

5.2.2 顔画像生成

この実験では、異なる被験者間での表情の入れ替え (swapping) と、2 種類の表情の補間 (interpolation) を行い、獲得された被験者特徴と表情特徴の分離度の評価と顔表情特徴の連続性の評価を行う。

まず、表情の入れ替えでは、ExpressionEncoder で抽出された表情特徴 z_e を異なる被験者で入れ替えた特徴を Decoder に入力し、画像生成を行う。被験者によらない表情特徴の獲得が行えている場合、表情特徴が異なる被験者間で入れ替わったとしても、表情や被験者の特徴が崩れることなく画像生成が行えるはずである。図 4 (a), (b) はそれぞれ学習データ、テストデータに対する表情入れ替えの結果を表している。最上段の画像は表情を埋め込む対象となる画像、左端の画像は埋め込む表情の画像である。また、それぞれの行は従来手法、提案手法 (繰り返し学習なし)、提案手法 (繰り返し学習あり) の生成結果を表しており、括弧の中に示した数字は入力画像のサイズである。まず

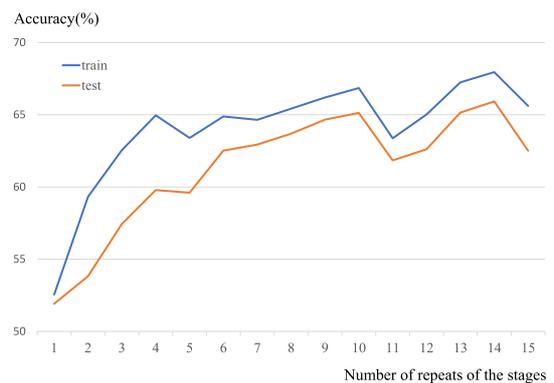


図 3 学習ステージ繰り返し回数による認識精度変化

Fig. 3 Change in recognition accuracy depending on the number of times the learning stage is repeated.

(a) の従来手法の結果に注目すると、生成画像がぼやけていたり、表情が崩れていることが確認できる。一方、提案手法では、生成画像は鮮明になり、被験者特徴が崩れることなく細かなシワの部分まで再現できていることが確認できる。また、繰り返し学習の有無で結果を比較すると、繰り返し学習を行った場合ではより画像が鮮明化されており、口や眉間のしわが生成画像に現れていることが分かる。この点から、提案手法に施された改良により、細かな表情の特徴が獲得できただけではなく、被験者特徴と表情特徴をうまく分離し、被験者によらない表情特徴として獲得できていることが確認された。(b) のテストデータに注目するとすべての結果について、被験者の特徴が変化してしまっていることが分かる。これは学習用の被験者 195 名では被験者特徴の多様性が不足しており、被験者特徴の学習が十分に行えなかったためと考えられる。しかし、表情についてはテストデータに対してもうまく入れ替えることができていた点から、モデルが被験者によらない表情特徴を獲得し、被験者特徴と表情特徴が分離されていることが確認できる。

次に、表情空間の連続性を検証するために、ソース画像とターゲット画像から計算された表情特徴 $z_{e_{src}}$ と $z_{e_{trg}}$ の間を補間することにより、中間の表情画像の生成を行う。

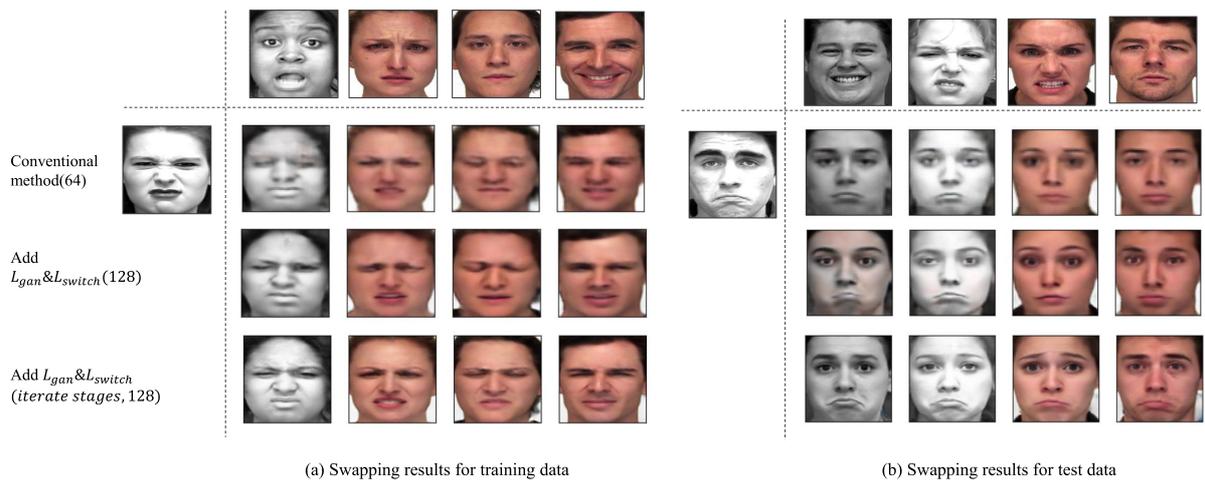


図 4 表情入れ替え画像生成結果

Fig. 4 (a) and (b) show the result of facial expression replacement for the training and test data. The first row shows the data sample for identity factor and leftmost image represents the data sample for facial expression factor. Each line is the result of swapping by each method.

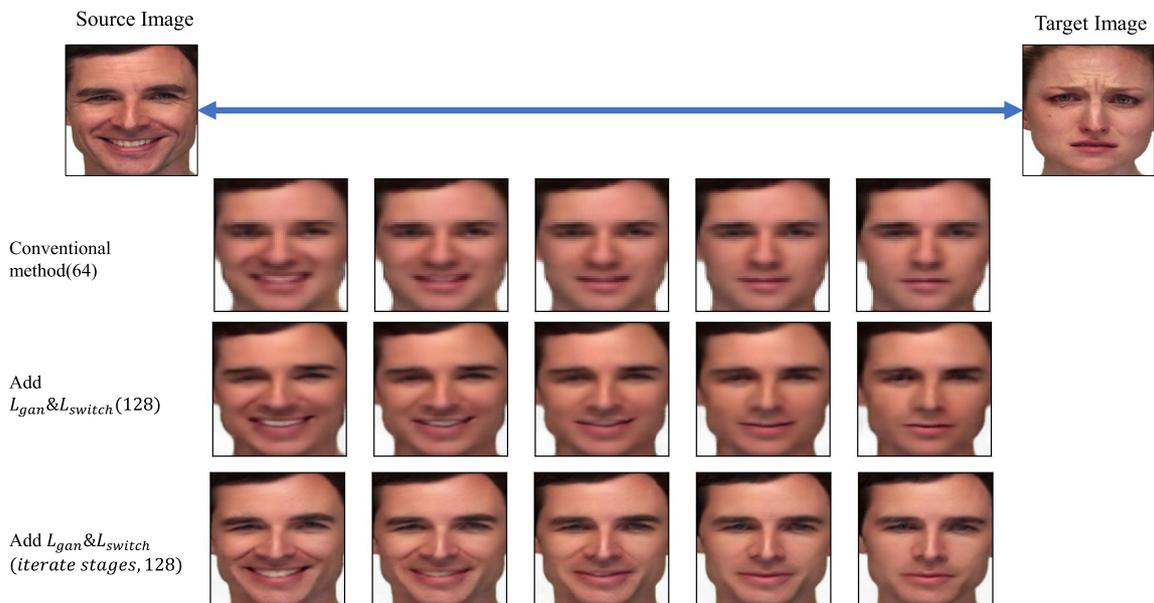


図 5 中間表情画像生成結果

Fig. 5 The result of intermediate facial expression image generation. This shows the result of gradually bringing the facial expression component of the source image closer to the facial expression component of the target image by each method.

もし、モデルが連続的な表情特徴の獲得が行われていれば、表情は $z_{e_{src}}$ で表される表情から $z_{e_{trg}}$ で表される表情へ、表情や被験者特徴が崩れることなく徐々に変化していくはずである。図 5 は表情補間の結果であり、最上段はソース画像とターゲット画像、2 段目以降は、それぞれの手法によって生成された補間画像を表している。まず、従来手法に着目すると、表情の補間はある程度行えているものの、表情の入れ替えの結果と同様に、画像はぼやけてしまっており、部分的に画像が崩れていることが確認できる。一方、提案手法では、中間の表情であっても鮮明な画像が生成さ

れていることが確認できる。また生成される表情も連続的に変化しており、加えて中間の生成画像は意味的にも中間的な表情になっていることが分かる。つまりソース画像の喜びとターゲット画像の悲しみの中間の画像は、悲しそうに笑っている表情になっている。このように、補間によって生成された画像の表情が鮮明さを保ったまま連続的に変化していることから、ExpressionEncoder の潜在変数として、連続的な表情特徴が獲得されたことが確認された。

6. まとめ

本稿では、顔のしわなど表情を構築するうえで重要である、より詳細な表情特徴を潜在変数として獲得することを目的として、文献 [9] で提案された手法の改良を行った。従来手法では、生成画像がぼやけることから細かな表情特徴が獲得できていないことが示唆されていた。そこで、提案手法では生成画像の鮮明化を目的とした損失関数を導入するとともに、表情特徴が被験者間で共通したものになることを目的とした損失関数を同時に導入することにより、被験者特徴と表情特徴が絡み合ったものになることを抑制しつつ、より詳細な表情特徴の獲得を可能にした。さらに学習ステップを従来手法の 1 巡のみ行う方式から複数回繰り返す方式に変更することで、2 つの Encoder が互いに影響を与え合うことを可能とし、被験者特徴と表情特徴をより分離した状態で獲得できるよう改良を行った。

実験では、ユークリッド距離を用いたクラスタリング手法による表情認識と顔画像生成（表情の入れ替え・補間）を行い、表情認識では従来手法から 10 % 以上（test : 11.62%, train : 15.97%）の精度向上を実現し、画像生成ではより鮮明な顔画像の生成とリアルな中間表情画像の生成を実現した。

今後は、提案手法によって獲得された連続的な表情の特徴を利用した新たな表情認識手法を検討したいと考える。

参考文献

- [1] Bengio, Y., Courville, A. and Vincent, P.: Representation learning: A review and new perspectives, *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol.35, No.8, pp.1798–1828 (2013).
- [2] Krizhevsky, A., Sutskever, I. and Hinton, G.: ImageNet classification with deep convolutional neural networks, *Advances in Neural Information Processing Systems*, pp.1097–1105 (2012).
- [3] Lopes, A.T., de Aguiar, E., De Souza, A.F. and Oliveira-Santos, T.: Facial expression recognition with convolutional neural networks: Coping with few data and the training sample order, *Pattern Recognition*, Vol.61, pp.610–628 (2017).
- [4] Yu, Z. and Zhang, C.: Image based static facial expression recognition with multiple deep network learning, *Proc. 2015 ACM on International Conference on Multimodal Interaction*, pp.435–442, ACM (2015).
- [5] Choi, Y., Choi, M., Kim, M., Ha, J.-W., Kim, S. and Choo, J.: Stargan: Unified generative adversarial networks for multi-domain image-to-image translation, *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp.8789–8797 (2018).
- [6] Ekman, P. and Friesen, W.V.: Constants across cultures in the face and emotion, *Journal of Personality and Social Psychology*, Vol.17, No.2, p.124 (1971).
- [7] Du, S., Tao, Y. and Martinez, A.M.: Compound facial expressions of emotion, *Proc. National Academy of Sciences*, pp.E1454–E1462 (2014).
- [8] Kostic, R., Alvarez, J.M., Recasens, A. and Lapedriza, A.: Emotion recognition in context, *IEEE Conference on Computer Vision and Pattern Recognition*, Vol.1 (2017).
- [9] Kanou, Y. and Nagao, T.: Separation of the Latent Representations into “Identity” and “Expression” without Emotional Labels, *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pp.1638–1644, IEEE (2020).
- [10] Kingma, D.P. and Welling, M.: Auto-encoding variational bayes, *The 2nd International Conference on Learning Representations*, Vol.abs/1312.6114 (2013).
- [11] Rezende, D.J., Mohamed, S. and Wierstra, D.: Stochastic backpropagation and variational inference in deep latent gaussian models, *Proc. 31st International Conference on Machine Learning*, Vol.32, No.2, pp.1278–1286 (2014).
- [12] Chen, X., Duan, Y., Houthoofd, R., Schulman, J., Sutskever, I. and Abbeel, P.: Infogan: Interpretable representation learning by information maximizing generative adversarial nets, *Advances in Neural Information Processing Systems*, pp.2172–2180 (2016).
- [13] Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S. and Lerchner, A.: beta-vaes: Learning basic visual concepts with a constrained variational framework (2017).
- [14] LeCun, Y. and Cortes, C.: MNIST handwritten digit database (2010) (online), available from <http://yann.lecun.com/exdb/mnist/>.
- [15] Liu, Z., Luo, P., Wang, X. and Tang, X.: Deep Learning Face Attributes in the Wild, *Proc. International Conference on Computer Vision (ICCV)* (2015).
- [16] Liu, Y., Wei, F., Shao, J., Sheng, L., Yan, J. and Wang, X.: Exploring disentangled feature representation beyond face identification, *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp.2080–2089 (2018).
- [17] Bouchacourt, D., Tomioka, R. and Nowozin, S.: Multi-level variational autoencoder: Learning disentangled representations from grouped observations, *32nd AAAI Conference on Artificial Intelligence* (2018).
- [18] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y.: Generative Adversarial Nets, *Advances in Neural Information Processing Systems*, Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. and Weinberger, K.Q. (Eds.), Vol.27, Curran Associates, Inc. (2014).
- [19] Aifanti, N., Papachristou, C. and Delopoulos, A.: The MUG facial expression database, *2010 11th International Workshop on Image Analysis for Multimedia Interactive Services*, pp.1–4, IEEE (2010).
- [20] Lucey, P., Cohn, J.F., Kanade, T., Saragih, J., Ambadar, Z. and Matthews, I.: The extended cohn-kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression, *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pp.94–101, IEEE (2010).
- [21] Livingstone, S.R., Peck, K. and Russo, F.A.: Ravdess: The ryerson audio-visual database of emotional speech and song, *Annual Meeting of the Canadian Society for Brain, Behaviour and Cognitive Science*, pp.205–211 (2012).
- [22] Ren, S., He, K., Girshick, R. and Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks, *Advances in Neural Information Processing Systems*, pp.91–99 (2015).

- [23] He, K., Zhang, X., Ren, S. and Sun, J.: Delving deep into rectifiers: Surpassing human-level performance on imagenet classification, *Proc. IEEE International Conference on Computer Vision*, pp.1026–1034 (2015).
- [24] Kingma, D.P. and Ba, J.: Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980 (2014).
- [25] Arthur, D. and Vassilvitskii, S.: k-means++: The advantages of careful seeding, *Proc. 18th Annual ACM-SIAM Symposium on Discrete Algorithms*, pp.1027–1035 (2007).
- [26] Kingma, D.P., Mohamed, S., Jimenez Rezende, D. and Welling, M.: Semi-supervised Learning with Deep Generative Models, *Advances in Neural Information Processing Systems*, Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. and Weinberger, K.Q. (Eds.), Vol.27, Curran Associates, Inc. (2014).



狩野 悌久

2019年横浜国立大学大学院環境情報学府情報環境専攻博士課程前期修了。現在、同博士課程後期在学中。画像処理を主として知能情報学の研究に従事。



長尾 智晴 (正会員)

1985年東京工業大学大学院総合理工学研究科博士課程後期中退。同年同大学助手。同大学助教授を経て、2001年横浜国立大学大学院環境情報研究院教授。工学博士。画像処理、進化計算法等の知能情報学の研究に従事。電子情報通信学会、人工知能学会、進化計算学会、IEEE等各会員。