Vol.2021-MPS-135 No.1
2021/10/18

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]　長尾　智晴[2]

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に対する提案手法の有効性を示す．

キーワード：弱教師あり学習，縺れを解いた特徴表現学習, 表情認識，画像生成

Weakly Supervised Learning for Acquisition of Continuous Facial Expression Features

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

1. はじめに

機械学習において，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．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 [1] や顔画像生成 [2] を行う研究では，Ekman の基本六感情 [3] などの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 [4] 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) [5] 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，獲得され

[1] 横浜国立大学大学院環境情報学府
[2] 横浜国立大学大学院環境情報研究院

ⓒ 2021 Information Processing Society of Japan

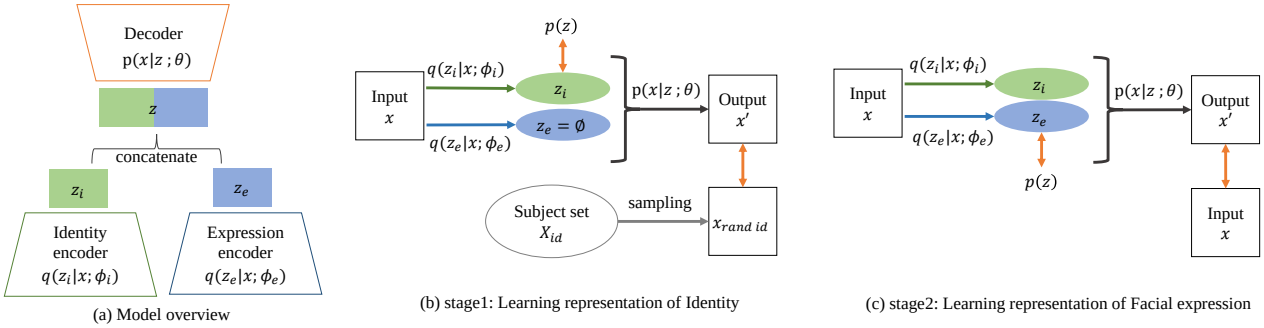

図 1　従来手法：モデル構造および学習ステージ

Fig. 1　(a) represents overview of the model and (b),(c) shows training steps. (b) shows the learning identity repsentation 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 $x_{rand_{id}}$ and output. (c) shows the learning facial expression repsentation stage. In this stage the IdentityEncoder is not update, and reconstruction error is calculated between input and output

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，表情認識と顔画像生成を実施し，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. 先行研究

先行研究 [4] は VAE を拡張し，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，被験者特徴の獲得と表情特徴の獲得をそれぞれ目的とした 2 段階の学習が行なわれる．

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 と Decoder 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．

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

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 z_e としてゼロベクトルが代入される．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．この損失関数により，モデルは同じ被験者のすべてのデータの組み合わ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．

次に，表情特徴の獲得を行うステージでは，IdentityEncoder のパラメータは固定し，ExpressionEncoder と Decoder 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 と同様に，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．

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

前段の学習により，すでに被験者特徴は 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．

3. 提案手法

先行研究では VAE の特性上，生成画像がぼやけており，

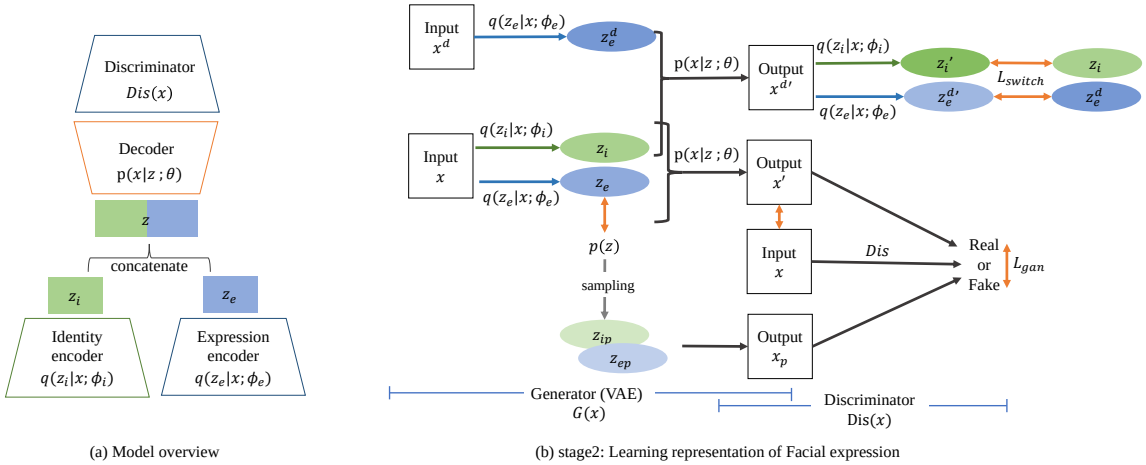

図 2　提案手法：モデル構造および表情特徴獲得ステージ

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

顔のしわなどの細かな特徴が捉えられていないことが示唆されていた．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で，より効果的な表情特徴の獲得を行う．

3.1　損失関数の改良

3.1.1　Adversarial Loss

Adversarial loss は GAN[6] 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 Discriminateor 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を行う．提案手法では，**図 2**(b) 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．

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

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．

3.1.2　Switch Loss

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．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．

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

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びそれぞれのエンコーダ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これらの損失関数と式（2）を組み合わせた目的関数 L_e' に従って学習を行う．

$$\mathcal{L}_e' = \mathcal{L}_e + \beta\mathcal{L}_{gan} + \gamma\mathcal{L}_{switch}, \qquad \beta, \gamma > 0 \tag{5}$$

ⓒ 2021 Information Processing Society of Japan

表 1　表情認識結果
Table 1　Results of facial expression recognition.

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

3.1.3　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ーの学習は，互いに影響を与えながら行われることが望ましい．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，また一巡のみ行われるために，第 1 ステージで学習される IdentityEcoder は ExpressionEncoder の学習結果を加味することができない．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し行うことで，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．

4.　実験

4.1　データセットおよび実験設定

この実験では MUG[7]，CK+[8]，RAVDESS[9] の 3 つのデータセットを組み合わせて利用する．実験ではこれのデータセットからそれぞれ 2 人~3 人の被験者を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，表情の多様性を確保するため，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，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を一致させるため，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抽出し作成する．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 29085 枚の画像，表情認識評価用のデータセットは 2037 枚（学習：1842 枚，テスト：195 枚）の画像を有するデータセットとして構成された．

表 2　エンコーダの構造
Table 2　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, fc_ σ	-	-	-	64

表 3　デコーダの構造
Table 3　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	$3 \times 64 \times 64$ ($3 \times 128 \times 128$)

エンコーダとデコーダの構造をそれぞれ**表 2**，**表 3** に示す．括弧で示した箇所は，入力サイズが $3 \times 128 \times 128$ の時に追加で用いるレイヤー，出力サイズを表している．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 らが提案した手法 [10] を用いて初期化され，最適化手法には Adam[11]（$\beta_1 = 0.9, \beta_2 = 0.999, \sigma = 1.0 \times 10^{-8}, lr = 0.0005$）を用いた．学習エポック数はそれぞれのステージで $100epoch$（繰り返し学習を行う場合は $50epoch$）とし，学習率はエポックごとに 0.95 を乗算し

ⓒ 2021 Information Processing Society of Japan

(a) Swapping results for training data　　(b) Swapping results for test data

図 3　表情入れ替え画像生成結果

Fig. 3　(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.

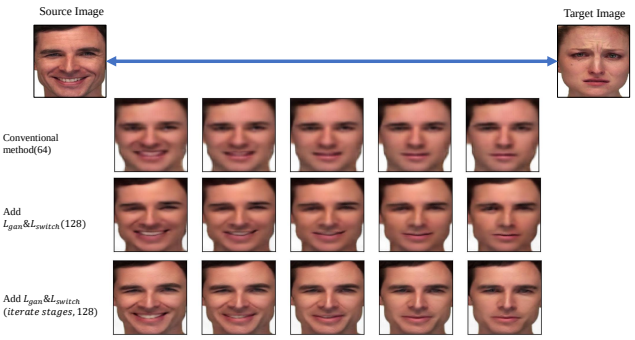

図 4　中間表情画像生成結果

Fig. 4　This figure 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.

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．

4.2　実験結果および考察

4.2.1　表情認識

この実験では，ExpressionEncoder の潜在空間を利用し，クラスタリング（k-meams+[12]）による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，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．比較手法は，VAE，C-VAE[13] の潜在空間を用いた場合と，VAE については，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により，被験者特徴を潜在変数から除く処理を加えた場合を実施した．

表 1 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 ともに教師なしで行われた結果である．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では，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，Swich 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伺える．さらに，提案手法では Adversarial loss を導入し細かな表情特徴の獲得を行っているため，入力サイズを 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言える．

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．

4.2.2 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と，2 種類の表情の補間を行い，獲得された被験者特徴と表情特徴の評価を行う．

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 3** 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の画像サイズである．まず (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，(c) の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ることが確認できる．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．

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．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 4** 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おり、表情も連続的に変化していることが確認できる．このことから，ExpressionEncoder の潜在変数として，連続的な表情特徴が獲得できていることが確認された．

5. まとめ

本稿では，[4] で提案された手法に対して，2 種類の損失関数を導入と学習ステップの改良を行うことにより，弱教師あり学習の枠組みでより効果的な表情特徴の獲得を行う手法を提案した．実験では，表情認識と顔画像生成に対して従来手法より高精度な結果を示した．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] 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).

[2] 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, *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8789–8797 (2018).

[3] 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).

[4] 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)*, IEEE, pp. 1638–1644 (2020).

[5] Kingma, D. P. and Welling, M.: Auto-encoding variational bayes, *The 2nd International Conference on Learning Representations*, Vol. abs/1312.6114 (2013).

[6] 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).

[7] Aifanti, N., Papachristou, C. and Delopoulos, A.: The MUG facial expression database, *Image analysis for multimedia interactive services, 2010 11th international workshop on*, IEEE, pp. 1–4 (2010).

[8] 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, *Computer Vision and Pattern Recognition Workshops, 2010 IEEE Computer Society Conference on*, IEEE, pp. 94–101 (2010).

[9] 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).

[10] He, K., Zhang, X., Ren, S. and Sun, J.: Delving deep into rectifiers: Surpassing human-level performance on imagenet classification, *Proceedings of the IEEE international conference on computer vision*, pp. 1026–1034 (2015).

[11] Kingma, D. P. and Ba, J.: Adam: A method for stochastic optimization, *arXiv preprint arXiv:1412.6980* (2014).

[12] Arthur, D. and Vassilvitskii, S.: K-means++: The advantages of careful seeding, *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 1027–1035 (2007).

[13] 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).

ⓒ 2021 Information Processing Society of Japan