横浜国立大学大学院環境情報学府
博士学位論文

機械学習における表情の連続性考慮と表情特徴獲得に関する研究

A Study on Handling Facial Expression Continuity and Acquiring Facial Expression Features in Machine Learning

情報環境専攻　情報学プログラム

狩野 悌久
Yoshihisa KANOU

請求学位　　博士（情報学）
責任指導教官　　長尾　智晴　教授

提出年月日　　令和 4 年 1 月 7 日
請求年度　　令和 3 年度 3 月修了

要約

近年の計算機の目覚ましい発展により，計算機の処理速度・計算可能量は飛躍的に増加している．それに伴い，今まで計算量がボトルネックとなっていた機械学習，特に深層学習に関連する研究は様々な分野で活発に行われ，多くのタスクに対して良好な結果を示している．画像処理の分野においては，Convolutional Neural Network（CNN）の登場により，分類や認識，生成を高精度に行うことが可能となり，様々なサービスや情報システムへの応用が期待されている．その中でも，表情を扱う領域は実社会への応用を期待されている分野の一つである．一般的に深層学習モデルの学習には大量のデータが必要であり，タスクに応じてそれぞれのデータに教師信号を与えるアノテーションと呼ばれる作業が必要になる．表情においては，データセット作成のコストの側面から，静止画像に対して ”怒り ”や ”喜び ”といった感情と紐づいた one-hot なラベルを利用してアノテーションを実施し，学習に利用することが多い．しかし，そのような静止画像を入力として学習を行った場合には，表情の時間的な連続性は考慮されず失われてしまう．また，one-hot なラベルがつけられたデータによってモデルを学習した場合には，モデルは表情を離散的なクラスとみなし，クラスごとの特徴量の離散化を図る方向で学習が行われる．しかし，表情は連続的に変化する事象であることから，このような離散的な扱いは避けるべきである．そこで，本論文では機械学習において表情の連続性を考慮する手法の検討を行う．

本論文ではまず，静止画像を対象とした表情認識モデルに対して時間方向の表情の連続性を考慮する機構を導入し，動画像の認識を効果的に行うための手法の提案を行う．静止画像を対象としたモデルをそのまま動画像に適用すると，隣接したフレーム間の微妙な表情変化であっても認識結果が大きく変動してしまう問題がある．そこで提案手法では，Attention 機構有した追加ネットワークをモデルに導入し，時間方向の特徴量に対して重みづけを行うことで，表情の時間方向の連続性を捉え，そのような認識結果の変動を抑制する．実験では動画像に対して表情認識を行い，動画像への提案手法の有効性の検証を行う．

次に，1 枚の画像から表情の連続性を保った表情特徴を抽出する手法の提案を行う．機械学習において表情を扱う場合，表情に対してアノテーションが行われることから，表情が離散的に扱われてしまう課題がある．提案手法では，表情に対するアノテーションを実施せずに，被験者情報（被験者 ID）を利用した 2 段階の学習により被験者特徴と表情特徴を分離した状態で潜在変数に獲得することで，表情本来の連続性を維持した表情特徴の獲得を可能にする．実験では，潜在空間の可視化や画像生成，表情認識によって提案手法によって獲得された特徴の評価を行う．

最後に，前章で提案した手法を改良し，1 枚の画像からしわなどのより詳細な表情の要素を捉えた表情特徴を抽出する手法の提案を行う．前章の手法では生成画像がぼやけて不鮮明であることから，細かな表情の特徴を抽出できていない可能性が示唆されていた．提案手法では，2 種類の損失関数の導入と学習方法の改良を行うことで，より細かな表情の特徴の抽出を行う．実験では，画像生成，表情認識のタスクを実施し，従来手法との比較を行うことで提案手法によって効果的な表情特徴が獲得されていることを示す．

Abstract

Because of the remarkable development of computers in recent years, the processing speed and computable amount of computers have increased dramatically. Along with this, research related to machine learning, especially deep learning, where the amount of calculation has been a bottleneck, has been actively conducted in various fields, and good results have been shown for many tasks. In the field of image processing, the advent of the Convolutional Neural Network (CNN) has made it possible to perform classification, recognition, and generation with high accuracy, and is expected to be applied to various services and information systems. Among them, the area dealing with facial expressions is one of the fields expected to be applied to applications. In general, learning a deep learning model requires a large amount of data, and a work called annotation that gives a teacher signal to each data according to the task. In terms of facial expressions, from the aspect of data set creation cost, annotate images using one-hot labels associated with emotions such as ”anger” and ”joy” and use them for learning. However, when learning is performed using such a still image as an input, the temporal continuity of facial expressions is not taken into consideration. In addition, when the model is trained with data labeled as one-hot, the model regards facial expressions as discrete classes, and training is performed in the direction of discretizing the features of each class. However, since facial expressions are continuously changing events, such discrete treatment should be avoided. Therefore, in this paper, we examine a method that considers the continuity of facial expressions in machine learning.

First, in this paper, we introduce a mechanism that considers the continuity of facial expressions in the time direction for a facial expression recognition model for images, and propose a method for effectively recognizing moving images. If a model for a still image is applied to a movie as it is, there is a problem that the recognition result fluctuates greatly even if there is a slight change in facial expression between adjacent frames. Therefore, in the proposed method, an additional network with an Attention mechanism is introduced into the model, and weighting is performed on the features of the time series. By doing so, the model can suppress fluctuations in the recognition result. In the experiment, facial expression recognition is performed on the moving image, and the effectiveness of the proposed method for the moving image is verified.

Next, we propose a method for extracting facial expression features that maintain facial expression continuity from a single image. When dealing with facial expressions in machine learning, there is a problem that facial expressions are treated discretely because of annotation for facial expressions. In the proposed method, facial expression features that maintain the original continuity of facial expressions are obtained by acquiring the subject features and facial expression features individually as a latent representation by two-step learning using subject information (subject ID) without annotating the facial expressions. In the experiment, the features acquired by the proposed method are evaluated by visualization of the latent space, image generation, and facial expression recognition.

Finally, we improve the method proposed in the previous chapter and propose a method to extract facial expression features that capture more detailed facial expression elements such as wrinkles from a

single image. Since the generated image is blurred and unclear by the method in the previous chapter, it was suggested that it may not be possible to extract the features of detailed facial expressions. In the proposed method, by introducing two types of loss functions and improving the learning method, more detailed facial expression features are extracted. In the experiment, the tasks of image generation and facial expression recognition are performed, and by comparing with the conventional method, it is shown that effective facial expression features are acquired by the proposed method.

目次

図目次

表目次

第1章　序論

1.1　背景と目的

近年の計算機の目覚ましい発展により，計算機の処理速度・計算可能量は飛躍的に増加している．それに伴い，今まで計算量がボトルネックとなっていた機械学習，特に深層学習に関連する研究は様々な分野で活発に行われ，多くのタスクに対して良好な結果を示している．画像処理の分野においては，Convolutional Neural Network（CNN）の登場により，分類や認識，生成を高精度に行うことが可能となり，様々なサービスや情報システムへの応用が期待されている．その中でも，表情を扱う領域は実社会への応用を期待されている分野の一つである．その理由としては，まずマンマシンコミュニケーションへの関心の高まりが挙げられる．今まで，マンマシンコミュニケーションとしては自然言語処理の分野が主流であり，チャットボットなどの自動応答システムとして応用がなされてきた．しかし，近年ではテキスト情報からだけでは取得できない感情情報を表情から抽出し，人間同士の会話で行われるような相手の表情から読み取れる感情や雰囲気によって応答を変えることを機械に行わせる，マルチモーダルな機械対話技術が注目を集めている．その他にも表情を扱う AI は，スマートフォンの普及から，写真加工や撮影，商品レコメンドを行うアプリケーションとしての需要も高まって来ている．

一般的に深層学習モデルの学習には大量のデータが必要であり，タスクに応じてそれぞれのデータに教師信号を与えるアノテーションと呼ばれる作業が必要になる．表情においては，データセット作成のコストの側面から，静止画像に対して”怒り”や”喜び”といった感情と紐づいた one-hot なラベルを利用してアノテーションを実施し，学習に利用することが多い．しかし，そのような静止画像を入力として学習を行った場合には，表情の時間的な連続性は考慮されず失われてしまう．また，one-hot なラベルがつけられたデータによってモデルを学習した場合には，モデルは表情を離散的なクラスとみなし，クラスごとの特徴量の離散化を図る方向で学習が行われる．しかし，表情は連続的に変化する事象であることから，このような離散的な扱いは避けるべきである．そこで，本論文では機械学習において表情の連続性を考慮する手法の検討を行う．本稿で考慮する表情の連続性とは以下の 2 点である

1. 時間方向に対する表情の連続性
2. 表情の連続性を有した表情特徴の獲得

1 点目は、離散的なクラスが付与された静止画像を利用して学習を行ったモデルに対して，時系列を考慮する機構を組み込むことで，時間的な連続性を踏まえた認識を実現し，認識精度を向上することが目的である．2 点目は，離散的なクラスを用いることなく，表情特徴の学習を行うことで，静止画像を利用して学習を行った場合であっても，表情特徴をモデル内部に表情の連続性を保った状態で獲得し，表情認識だけではなく画像生成やデータ拡張に獲得された特徴を利用することが目的である．

1.2 本論文の構成

本論文の構成は次のとおりである．まず第 2 章では，機械学習で表情を取り扱う際に利用されるアノテーション手法について紹介するとともに，本研究と関連の深い従来手法について述べる．第 3 章では，従来の表情認識で失われてしまう時間方向の表情の連続性を考慮することを目的として，静止画像を対象とした表情認識モデルを動画像に効果的に適用する手法を提案する．ここでは，動画像に対する表情認識を実験として実施し，提案手法の有効性を検証する．第 4 章では，表情の連続性を有した表情特徴の獲得を目的として，アノテーションによって与えられる表情に関する情報を用いることなく，モデル内部に連続的な表情特徴を行う手法を提案する．提案手法では，VAE の枠組みを拡張し，2 つの Encoder に異なる機能を与える学習を行うことで，それぞれの Encoder の潜在変数として被験者特徴と表情特徴を分離して獲得する．第 5 章では，第 4 章で提案した手法を改良し，より詳細な表情の特徴をモデル内部の特徴表現として獲得する手法の提案を行う．ここでは，表情認識と画像生成のタスクを実施し，提案手法と先行手法，従来手法を比較することで，提案手法が最も効果的な表情特徴を獲得できていることを検証する．最後に，第 6 章で本論文のまとめとして，本論文で得られた成果と今後の課題について述べる．

第2章　関連研究

本章では，まず表情認識とそれに用いられるアノテーション手法について述べ，機械学習の領域における一般的な表情の扱いについて説明を行う．続いて，本稿で提案する手法と関連の深い手法と研究領域について紹介する．

2.1　表情に対するアノテーション手法と表情認識

表情の符号化には脳科学・心理学・生物学などの側面から様々な手法が存在し，目的や特性に応じて使い分けられることが多い．ここでは，表情認識を行う際に用いられる代表的なアノテーション手法を紹介し，それぞれの特性やメリット・デメリットについて述べる．

2.1.1　基本六感情

Ekman の基本六感情 [3]は，機械学習に用いられる最も一般的な表情の符号化手法である．これは，人間の感情を基本的な 6 つの感情（Anger・Fear・Sadness・Surprise・Disgust・Happiness）に分ける理論である．表情認識を行う際には，これらの感情に標準状態を表す Neutral を加えた 7 つの感情をクラスとして表情に対して割り当てることでアノテーションを行い，分類タスクとして認識が行われることが多い [4] [5]．また，より細かな感情の認識を目的として，基本六感情を複数組み合わせ，Happily Surprised などの 21 種の感情を設定し認識を目的とした研究 [6]や，写真の背景や一緒に写っている人間の情報を加味することで 26 種類の感情の認識を試みる研究 [7]も行われている．

このような基本六感情を用いる方法は，専門的な知識を用いることなくアノテーションを行うことが可能であり，作業者を集めやすいことから，大規模データセットの構築が比較的容易である点がメリットとして挙げられる．一方で，素人によって判断が行われるため，与えられたラベルが主観的なものになりやすく，教師信号そのものが曖昧性を持つ可能性がある．また，これらの手法は，感情に対してクラスを割り当て，その感情と表情との間で対応をとる手法であることから，連続的な表情変化を離散的に扱うアノテーション方法であると言える．

2.1.2　VAD emotional state model

VAD emotional state model [8]は人間の感情を連続的な値として符号化する手法として知られている．この手法は人間の感情の基礎要素として，Valence（感情価），Arousal（覚醒度），Dominance（支配度）の 3 つの因子を定義し，それらの組み合わせによって人間の複雑な感情を表す手法である．表情認識においては，連続的に評価した感情と表情と対応付けを行うため，表情に対して連続性を考慮したアノテーションを行うことが可能である．一方で，3 つの要素を正確に測定し，判断するためには専門的な知識が必要となるため，アノテーション作業のコストが非常に高く，大規模データセットの構築が難しいという課題がある．

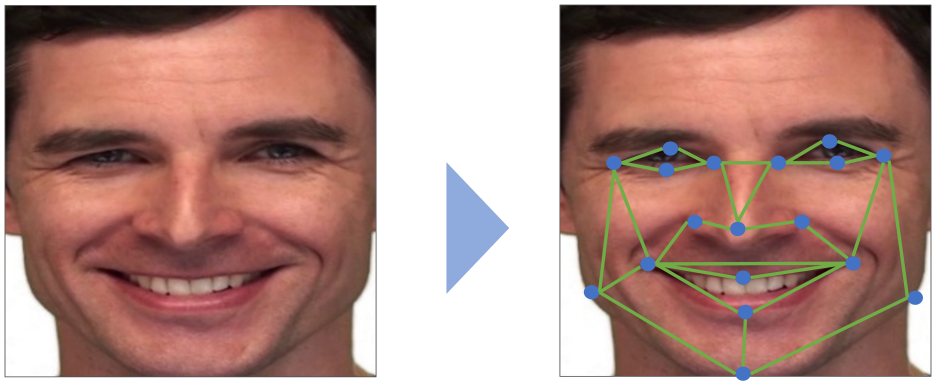

図 2.1: 顔特徴点

2.1.3 Facial Action Cording System (FACS)

Facial Action Cording System (FACS) [9]は前述した 2 つのように感情と表情を結び付けることで表情の符号化を行う手法とは異なり，表情を視覚的に検知できる顔の筋肉の動きとしてとらえ，表情を直接符号化する手法である．FACS は Action Unit(AU) と呼ばれる顔の筋肉の動作単位を組み合わせることで，人間の顔に現れる多様な表情を記述する．機械学習では顔画像からのいくつかの AU を検出するタスクとして行われることが多く，様々な手法が提案されている．AU 検出の主流のアプローチは，2.1 に示すような顔特徴点を利用し，特徴点の動きやその周辺のパッチ画像を入力とする方法である [10] [11]．一方で，Andres らは実世界での顔特徴点検出の難しさから，特徴点を利用せず顔画像全体を入力として高精度に AU の検出を行う手法の提案を行った [12]．

FACS は多くのパターンの表情を，AU という記述子の組み合わせによって正確に表現できるため，表情の解析や認識において有効である．一方で，顔画像からどの AU が現れているのかを判断するには非常に専門的な知識が必要となるため，深層学習を利用する場合には，大規模なデータセットの構築が課題となっている．

2.2 Attention 機構

Attention 機構とは，自然言語処理や画像処理を目的とした深層学習モデルに用いられる仕組みであり，中間層の値に対して動的に重みづけを行うことで，抽出された要素ごとの関連性や注目個所を学習することを目的とした手法である．

自然言語処理の分野では，2.2 に示すように時間方向に対して注目個所を決定するために利用されることが多く，Seq2Seq [13]と呼ばれる代表的な手法の改良手法として導入され，大きく精度が向上したことで注目を集めた [14]．現在では，BERT [15]を始めとする多くの手法に導入され，自然言語処理の様々なタスクにおいて state-of-the-art を実現している．

画像処理の分野では，空間方向・チャネル方向に対して注目個所を決定するために利用される．まず，チャネル方向の Attention とは，2.3 に示すように，中間層で得られる特徴量マップのどのチャネルに着目するかということである．一般に深層学習では，チャネルごとに異なる特徴（例え

ば同じ層のあるチャネルでは横方向のエッジ特徴，またあるチャネルでは縦方向のエッジ特徴など）を抽出していると考えることができる．そのため，チャネルごとに重みづけを行う操作は，抽出された特徴の中でどの特徴が有効であり，どの特徴が不要であるかを決定する操作と捉えることができる．この操作により，不要な特徴が下層に伝播することが抑制され，有用な特徴のみ利用できるようになるため，精度の向上が期待できる．

次に，空間方向とは画像中の位置を差し，Attention 機構は画像中のどの領域に注目するかを決定する．代表的な手法である Residual Attention Network [2]を例に説明すると，2.4 に示すように中間層の値である特徴量マップのどの位置の特徴の影響を強めるかを決定する Attention マスクを枝分かれしたネットワークによって生成し，この Attention マスクと特徴マップを掛け合わせることで，重み付きの特徴量マップを得る．この重み付き特徴量マップは，認識であれば認識に重要な領域，具体的には認識対象の特徴が現れる箇所の特徴が強まり，逆に認識に不必要な背景情報などを弱まっていることが期待される．これにより，単純なネットワークに比べて認識精度の向上が見込まれる．

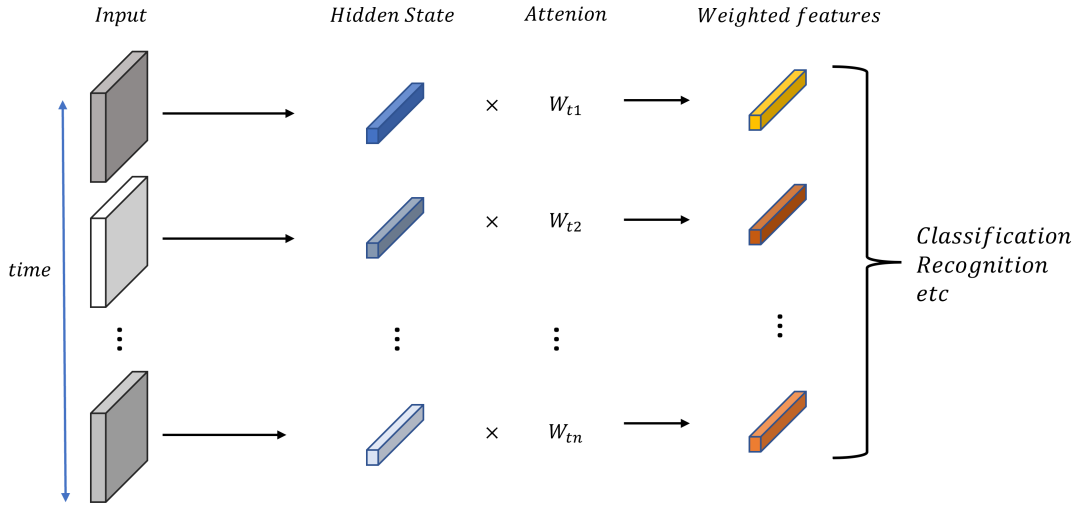

図 2.2: 時間方向の Attention

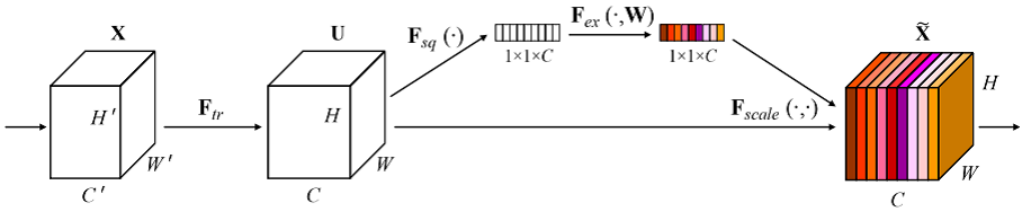

図 2.3: チャネル方向の Attention（画像は文献 [1]より引用）

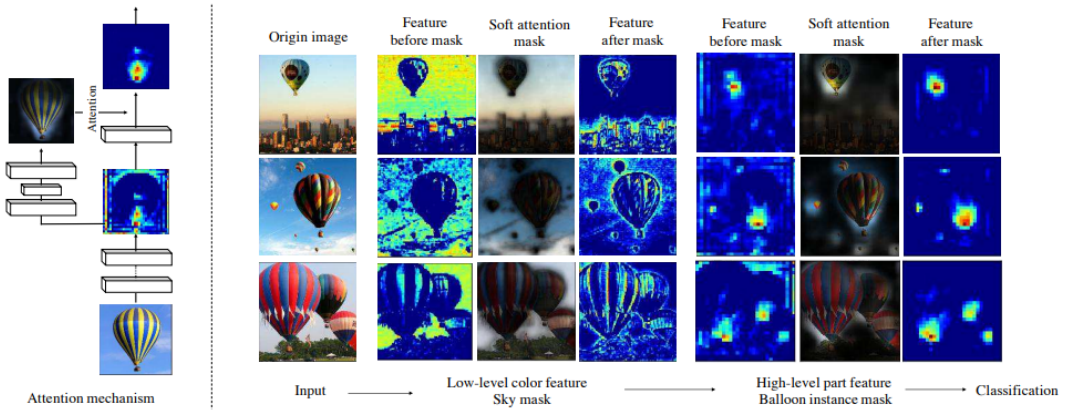

図 2.4: 空間方向の Attention（画像は文献 [2] より引用）

2.3 Variational Autoencoder

Variational Autoencoder（VAE）[16] [17]は表現力の高い潜在変数を獲得できる確率モデルとして知られている．一般的に Autoencoder は Encoder（$q_\phi(\boldsymbol{z}|\boldsymbol{x})$）と Decoder（$p_\theta(\boldsymbol{x}'|\boldsymbol{z})$）から構成されており，Encoder は入力データ x を潜在変数 z に変換し，Decoder は潜在変数 z を出力 x' に変換する．この時，入力データ x と出力 x' の間で再構築誤差をとることで，潜在変数 z には入力データを構築する情報が埋め込まれることが期待されるため，Autoencoder は教師なしで次元圧縮や表現力の高い特徴量を獲得するために多く用いられる手法である．VAE が通常の Autoencoder と異なる点は，潜在変数 z に対して確率分布を仮定し，潜在変数に制約を加えている点にある．多くの場合は，潜在変数の事前分布として正規分布 $\mathcal{N}(0, I)$ を仮定し，式 (2.1) に示すように Encoder を用いて平均ベクトル μ と分散ベクトル σ を求める．

$$\boldsymbol{z} \sim q_\phi(\boldsymbol{z}|\boldsymbol{x}) = \mathcal{N}(\boldsymbol{z}; \mu_\phi(x), \sigma_\phi(x)) \tag{2.1}$$

生成モデルの目的は，データ X の分布 $p(X)$ の推定を行うことであることから，VAE についても同様に，$p(X)$ の尤度を最大化する VAE のパラメータ ϕ, θ を学習によって決定する．実際は周辺対数尤度 $log p_{\phi,\theta}(X)$（ϕ, θ は VAE のパラメータ）の最大化をターゲットとして学習が行われる．しかし，周辺対数尤度を直接最大化しようとすると，積分の扱いが困難なため，誤差逆伝搬による学習を行うことができない．そこで，式 (2.2) に示すように変分下限 $L(\boldsymbol{x}, \boldsymbol{z})$ を最大化することで，周辺対数尤度を最大化を行い，パラメータの決定を行う．

$$\begin{aligned} \log p(\boldsymbol{x}) &\geq L(\boldsymbol{x}, \boldsymbol{z}) \\ &= -D_{KL}(q_\phi(\boldsymbol{z}|\boldsymbol{x}) \| p(\boldsymbol{z})) + E_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})] \end{aligned} \tag{2.2}$$

式 (2.2) の第 1 項は Encoder の潜在変数の分布 $p_\theta(\boldsymbol{x}|\boldsymbol{z})$ と事前分布として仮定した正規分布との差異を表し，第 2 項は入力データと出力データ間の再構築誤差を表している．ここで，変分下限の最大化を行う際の問題点として，式 (2.2) の第 1 項に式 (2.1) で示したような潜在変数 z のサンプリングが含まれている点である．このサンプリングは離散的な処理であるため，Encoder のパラメータ ϕ について微分を実施することができず，このままでは勾配降下法によるパラメータの更新が行えない．そこで，VAE では Reparameterization Trick と呼ばれる手法を用い，離散的なサンプリング

を行わずに潜在変数の決定を行う．Reparameterization Trick では，$z \sim \mathcal{N}(z; \mu_\phi(x), \sigma_\phi(x))$ を直接扱うのではなく，式 (2.3) に示すように，正規分布に従うノイズ ϵ を用いることで，確率変数が学習パラメータに含まれることを避け，勾配降下法による学習を可能にしている．

$$z = \mu_\phi(x) + \sigma_\theta(x) \odot \epsilon, \quad \epsilon \sim \mathcal{N}(0, I) \tag{2.3}$$

VAE の特性としては，まず潜在変数に制約をかけ正規分布を事前分布として仮定していることから，各々のデータの潜在変数が離散的になることが抑制され，潜在空間（潜在変数の属する N 次元空間）が連続的になりやすい．また学習過程で正規分布に従うノイズ ϵ を利用していることから，潜在空間において隣接した潜在変数同士は似通った表現になりやすいという性質がある．

2.4 縺れを解いた特徴表現の獲得

一般に機械学習で獲得された特徴表現は，様々な要素が複雑に絡み合ったものになるため，どの要素が特徴量のどの箇所と対応しているのかを人間が判断することは困難である．Disentangled representation learning（縺れを解いた表現学習）はそのような問題を解決し，特徴量を人間にとって意味的に解釈可能な状態で獲得することを目的とした学習方法である．このような表現の学習はコンピュータビジョンの分野においては，画像認識のタスクに対して有効な特徴を選別して認識に利用することを目的として利用されたり，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ことが多い．また，縺れを解いた表現を獲得するためには様々な手法が提案されており，学習データの観点から見ると，付加情報を利用せずに教師なしで行われる手法やラベルなどの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手法が存在する．

まず，教師なし学習の枠組みで解釈可能な表現を学習する手法としては，infoGAN [18]や β-VAE [19]がよく知られている．infoGAN は，ノイズから画像生成を行うタスクにおいて，教師なしで生成画像をコントロールするために提案された手法であり，入力となる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 に分解することを目的とする．この手法では，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 に解釈可能な表現の獲得を実現する．β-VAE は，VAE の近似事後分布を事前分布にどの程度近づけるか決定する正則化項 β を導入し，潜在空間に対する制約を制御することで，縺れを解いた表現を獲得する手法である．これらの手法は，Mnisit [20]などの比較的単純なデータに対して良好に縺れを解いた特徴表現の獲得を実現できる．一方で，CelebA [21]などの様々な要素が存在する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ことが課題として挙げられる．

そこで，教師信号を利用し，より複雑なデータに対して，縺れを解いた特徴表現の獲得を行う手法の研究が行われている．DR-GAN [22]は顔の向きの変化に頑健な顔認証のための特徴獲得を目的とした手法である．DR-GAN では Generative Adversarial Network(GAN) [23] の枠組みを利用し，被験者を維持したまま任意の顔の向きの画像を生成することで，顔の向きに影響されない被験者の特徴を獲得している．この手法では顔の向き情報 c が教師として与えられ，この c に顔画像における顔の向きに関する影響を内包させる事で，特徴量として顔の向きに影響されない被験者特徴の獲得を実現している．教師あり学習で縺れを解いた特徴表現の獲得を行う手法は，タスクに対して有益な特徴量獲得において強力であるが，データへのラベリングのコストが高くなるため，実世界の問題に対して適用が難しいという課題がある．

そこで，直接の教師信号ではなく付加情報を利用した弱教師あり学習の枠組みで学習を行うことでラベリングのコストを抑えつつ，複雑なデータに対して縺れを解いた特徴表現の獲得を目的とし

た研究が行われている．Liu らは付加情報として被験者情報を用いることで，顔画像から帽子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 [24]を提案した．また，Dian 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) に分けた状態で獲得する ML-VAE [25]を提案している．これらの手法では，比較的与えやすい情報を付加情報として利用し，付加情報を頼りにして学習を行える点から，複雑なデータであっても比較的良好に縺れを解いた特徴表現の獲得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ことから，獲得された特徴と画像に現れる各属性の対応をとる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のみであり，その連続性は考慮されていない．

2.5 まとめ

本章では，表情に対するアノテーション手法に触れ，機械学習における一般的な表情の取り扱いとその課題について説明を行うとともに，提案手法と関わりの深い研究領域とその領域における先行研究について述べた．

次章からは，2.1 で述べた，表情の符号化の課題である,「表情の連続性の欠如」「教師信号の曖昧性」「データ作成のコスト」の面から，機械学習において，表情を扱う際にこれらの課題を解決するための手法の提案を行う．

第3章　静止画像を対象とした表情認識モデルの動画像への効果的な適用手法の提案

3.1　はじめに

第 2 章で紹介したアノテーション手法を用いて，様々な顔画像データセットが構築され，機械学習に利用されている．しかし，大規模なデータセットとして構築されているものに注目すると，1 枚の静止画像に対して基本六感情などの離散的なラベルを用いてアノテーションされたものがほとんどであり，動画像に対して連続的な情報を与えたものはほとんど存在しない．そのため，深層学習モデルを利用した表情認識のタスクでは，静止画像を対象として離散的なラベルを推定することが主流となっている．しかし，そのような離散的なラベルが付与されたデータを用いて学習を行った認識器では，表情の時間的な変化を捉えることができないため，動画像に適用した際に，隣接フレームの微妙な表情変化であっても，認識結果が大きく変動してしまい，認識として不自然な結果になってしまう．

そこで本章では，時間方向の連続性を考慮するために Attention 機構を有したネットワークを静止画像を対象としたモデルに導入することで，動画像を認識対象としたモデルに拡張する手法を提案する．この手法では，表情の連続性として，時間方向に対する表情の連続性について取り扱う．実験では，提案手法を動画像に適用した際の認識結果について示すことで，提案手法の動画像に対する有効性の検証を行う．

3.2　静止画像モデルの動画像への拡張

提案モデルの概要を 3.1 に示す．提案モデルは，静止画像を認識対象としたモデルと時間方向の連続性を考慮するための追加ネットワークからなる．これらのモデル・ネットワークはそれぞれ別々に学習が行われ，その機能を獲得する．そのため，静止画像を認識対象としたモデルは学習済みの既存モデルをそのまま利用することができ，本手法は様々なモデルに対して，時間方向の連続性を考慮する仕組みを組み込むことを可能にする．

3.2.1　静止画像を対象としたモデルの学習

この学習フェーズでは，大量に収集可能な静止画像に対して離散的なラベルを用いてアノテーションされたデータセットを利用して，Neural Network（NN）をベースとした認識モデルの学習を行う．この学習の目的は，大量のデータセットを利用して，表情認識のタスクに対して有益な特徴を抽出する下位レイヤ（$dl_\theta()$）とその特徴から認識を行う上位レイヤ（$ul_\phi()$）の学習を行うことである．モデルの学習は，NN によって分類タスクを行う際に広く利用される，Softmax Cross

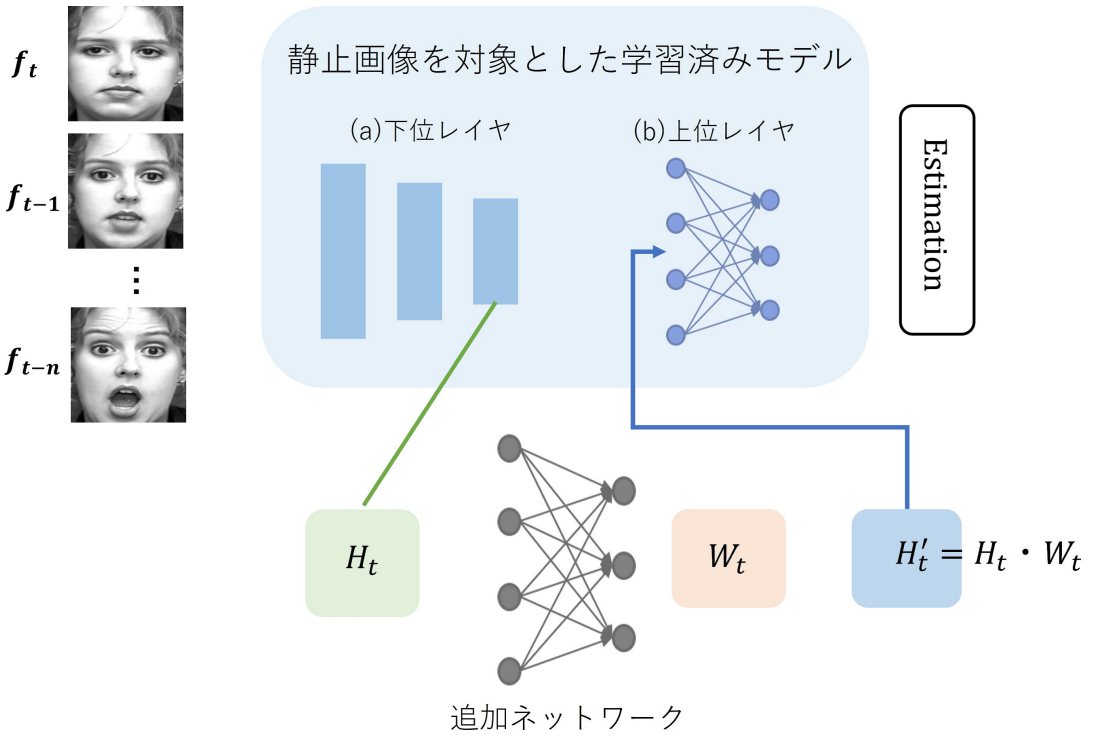

図 3.1: 提案モデル概要

Entropy（式 3.1）を誤差関数として用いて行われる．

$$\begin{aligned} L_{sce} &= -\sum_i t_i log(S(y_i)) \\ S(y_i) &= \frac{e^{y_i}}{\sum e^{y_i}} \\ y &= ul_\phi(dl_\theta(x)) \end{aligned} \tag{3.1}$$

$S(y_i)$ は softmax 関数を，x は入力，y はモデルの出力値を表している．また，t は教師信号として与えられた one-hot なラベル情報であり，i はベクトルデータの i 番目の要素である．

このフェーズの学習で獲得されたモデルのパラメータは，学習終了時に固定され，次フェーズの学習ではパラメータの更新は行われない．

3.2.2 時間方向の連続性を考慮する追加ネットワークの学習

このフェーズでは，前フェーズで学習を行った静止画像を対象とした表情認識モデルを効果的に動画像へ適用することを目的とした，時間方向に対して重みづけを行う Attention 機構を有した追加ネットワークを導入し，学習を行う．図 3.2 に追加ネットワークを用いて時系列方向に重みづけを行う流れを示す．まず，動画像の時間幅 n で切り出した隣接フレーム（$f_{t-n}, ..., f_{t-1}, f_t$）をそれぞれ学習済みのモデルの下位レイヤに入力し，特徴量 $(h_{t-n}, ..., h_{t-1}, h_t)$ を抽出する．次に，それらの時系列に並んだ特徴量を concatinate した特徴 $H_t = (h_{t-n}, ..., h_{t-1}, h_t)$ を追加ネットワークの入力とし，この時系列特徴量それぞれに対する重み係数 $W_t = (w_{t-n}, ..., w_{t-1}, w_t)$ の算出を行う．その後，H_t と W_t は乗算され重み付き特徴 H'_t が算出される．最後に，H'_t を学習済みモデルの上位レイヤに入力

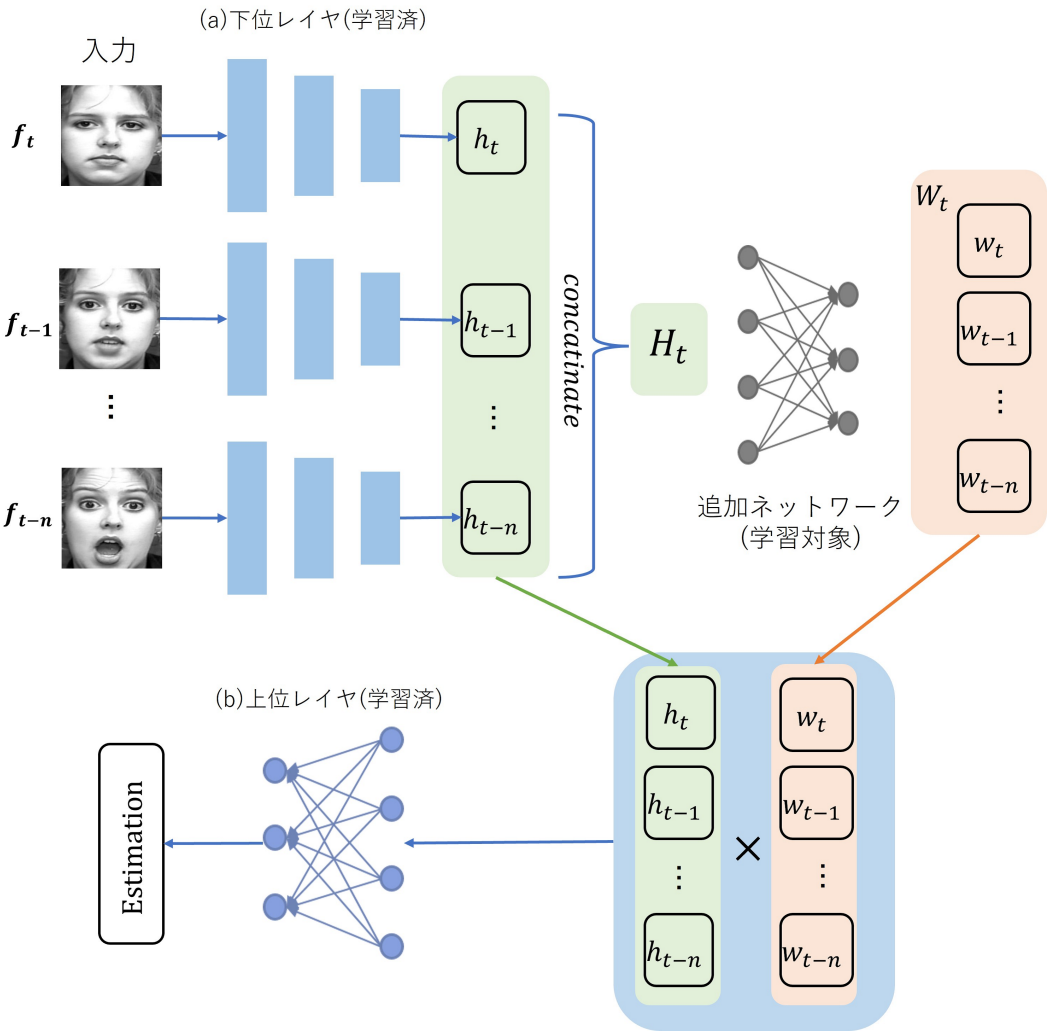

図 3.2: 追加ネットワークによる時系列特徴量の重みづけ

し，表情認識を行う．このようにすることでモデルは時間幅 n の時系列特徴を考慮することができるようになるとともに，時間幅の中で有益な特徴の影響を強め，ノイズとなる不要な特徴の影響を抑えることが可能になる．

ここで，追加ネットワークの目的は表情の時系列方向の連続性を考慮し，隣接フレーム間で認識結果が大きく変動することを抑制することである．そこで，式 3.2 に示すような誤差関数 L_{ex} を設定し，ネットワークの学習を行う．

$$
\begin{aligned}
L_{ex} &= \frac{1}{n}\sum_i (y_i - y'_i)^2 \\
y &= ul_\phi(H'_t) \\
y' &= ul_\phi(H'_{t-1})
\end{aligned}
\tag{3.2}
$$

y と y' は，それぞれ同じ動画内で 1 フレームだけずらした重み付き特徴 H'_t と H'_{t-1} を上位レイヤに入力したときの出力値を表している．また，L_{ex} は y と y' の平均二乗誤差である．これにより，追

加ネットワークの学習は，隣接フレーム間では特徴が大きく変動しないよう時系列特徴に対して重みづけを行うように行われるため，結果として認識結果の変動を抑えることを実現する．また，この誤差関数は，表情に対するアノテーションを必要せずに学習を行うことが可能である．つまり，提案手法は，静止画像を対象としたモデルを学習した際に利用した離散的な表情ラベルのみの利用でありながら，表情の時間方向の連続性をモデルが扱うことを可能にしている．

3.3 動画像に対する表情認識実験

本実験では，提案手法の動画像に対する表情認識への有効性の検証を行う．ここで,「有効性」とは動画像に対する認識結果が隣接フレーム間で大きく変動することなく，結果として自然な認識を実現する事である．

3.3.1 実験設定

データセットと前処理

本実験では，学習用のデータセットとして 2 種類のデータセットを利用する．まず，表情認識モデル学習用のデータセットとして，表情認識のベンチマークとして一般的に用いられるデータセットである FER2013 [26]を利用する．FER2013 は，画像サイズ 48×48 のグレースケール画像，合計 35885 枚から構成されており，それぞれのデータに対して基本六感情＋ Neutral の 7 クラスの one-hot なラベルが付けられている．

追加ネットワークには，表情の動画像データセットである，MUG [27]を利用する．MUG は 19fps の 3 6 秒の動画像から構成されており，被験者は Neutral な表情から基本六感情のうち一つを表現し Neutral な表情に戻る演技を行う．今回の実験では，公開されている 52 名の被験者の 862 本の動画を利用した．

入力前の処理として，MUG については上半身動画像であるため，顔検出を行い，顔領域の切り出しを行う．また，すべてのデータは 64 × 64 にリサイズしたのち，モデルに入力される．

モデル構造と学習方法

本実験で利用する静止画像を対象としたモデルと追加ネットワークの構造を表 3.1，表 3.2 に示す．静止画像を対象とした表情認識モデルは，ResNet [28]で提案された Residual Block を有するモデルであり，fc1 までを下位レイヤ，fc2 を上位レイヤとして扱う．ResNet は画像認識の様々なタスクにおいて良好な結果を示すことが知られているため，今回の実験に用いるモデルとして採用した．追加ネットワークは，2 層の全結合層からなるニューラルネットワークである．今回の実験では，時系列を考慮する時間幅を $n = 10$ と設定したため，追加ネットワークの入力は，下位レイヤの出力の 10 倍である 640 次元の時系列特徴ベクトルとし，出力は時間幅と同様に 10 になっている．

それぞれのネットワークの重みパラメータは，He らが提案した手法 [29]によって初期したのちに学習を行った．学習には Momentum SGD（モーメンタム $\mu = 0.9$）を利用し，学習率の初期値は $lr = 0.1$ とした．また，ミニバッチサイズは 100 に設定した．静止画像モデルの学習時はこの設定で 200 エポックの学習を行い，100 エポック目と 150 エポック目で学習率に対して 0.1 を乗算し減衰させた．また，データの水増しとして，クロッピング，鏡面反転，減色処理，ガンマ補正，ノイ

ズの付与，オクルージョン処理をランダムで行った．追加ネットワークの学習時にはエポック数を 50 とし，25 エポック目と 40 エポック目で学習率に対して 0.1 を乗算し減衰させた．これらのエポック数，減衰の仕方は実験的に決定したものである．

表 3.1: 静止画像を対象とした表情認識モデルの構造

Layer	Type	Output	Detail (kernel size, channels)
Input	Image data	$1 \times 64 \times 64$	-
lower layer	Conv 1	$64 \times 32 \times 32$	$3 \times 3, 64$
	Residual Block 1	$128 \times 32 \times 32$	$\begin{pmatrix} 1 \times 1, 64 \\ 3 \times 3, 64 \\ 1 \times 1, 128 \end{pmatrix} \times 2$
	Residual Block 2	$256 \times 16 \times 16$	$\begin{pmatrix} 1 \times 1, 128 \\ 3 \times 3, 64 \\ 1 \times 1, 256 \end{pmatrix} \times 2$
	Residual Block 3	$512 \times 8 \times 8$	$\begin{pmatrix} 1 \times 1, 256 \\ 3 \times 3, 128 \\ 1 \times 1, 512 \end{pmatrix} \times 2$
	Residual Block 4	$1024 \times 4 \times 4$	$\begin{pmatrix} 1 \times 1, 512 \\ 3 \times 3, 256 \\ 1 \times 1, 1024 \end{pmatrix} \times 2$
	global average pooling	$1024 \times 1 \times 1$	-
	fc 1	64	-
upper layer	fc 2	7	-

表 3.2: 追加ネットワークの構造

Layer	Type	Output	Detail(kernel size, channels)
Input	Time series features	640	-
Extra Network	fc 1	256	-
	fc 2	10	-

3.3.2 実験結果

図 3.3 は，追加ネットワークの学習に用いた MUG データセットの未知の被験者に対して，3 つの方法で表情認識を行った結果を時系列で表したものである．結果は，(a) は静止画像モデルによってフレームを 1 枚ずつ認識した結果，(b) は (a) の結果に対して 10 フレームで移動平均をとった結果，(c) は提案手法による認識結果を表している．また，適用したデータは表情が Neutral から Anger になり，再び Neutral に戻る動画像である．まず (a) の静止画像モデルで認識した結果に着目すると，認識結果が大きく変動していることが確認できる．また，(b) の結果から移動平均を取ることである程度のブレは抑制することができるが，局所的に現れる大きな認識結果のブレによって認識結果

が乱れたり，異なる表情クラスの出力が大きくなってしまうことが伺える．一方，提案手法の結果では，非常になめらかな認識結果となっており，局所的な大きなブレにも影響されずに認識を行えていることが分かる．これは，追加ネットワークが各フレームから抽出された時間幅を持った特徴量に対して，局所的な認識のブレにつながる特徴の影響を抑えるよう重みづけを行えてい，認識結果のブレを抑制しているためであると考えられる．

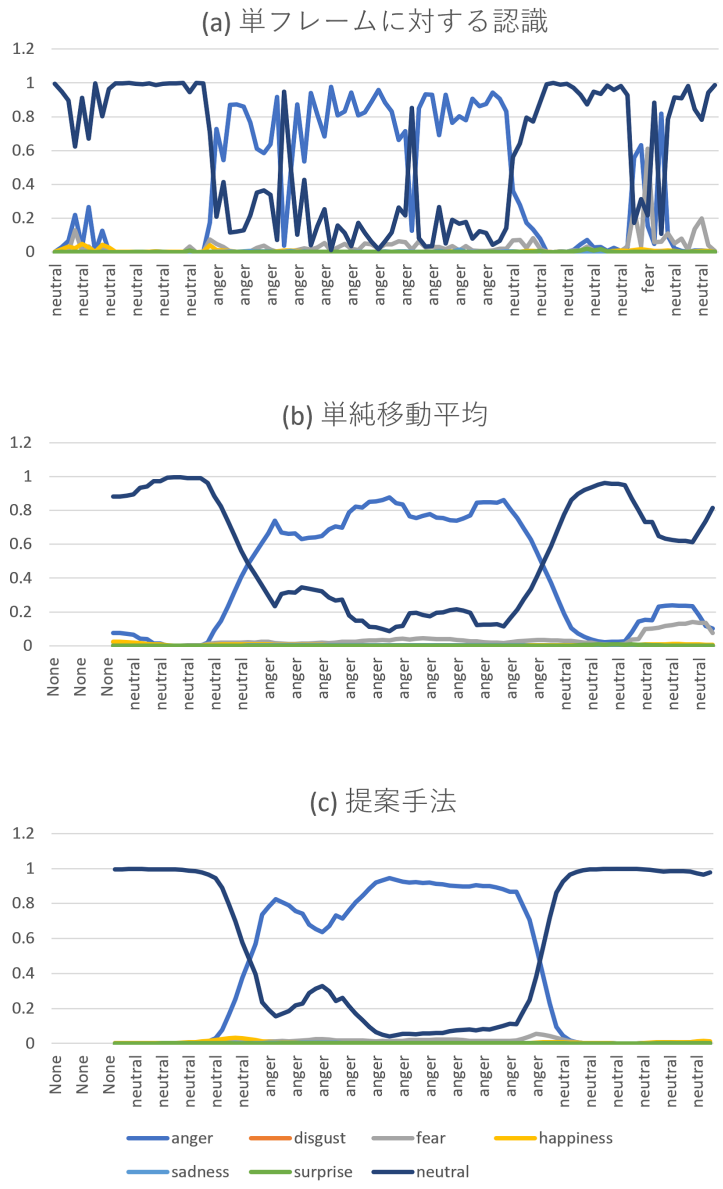

図 3.3: 動画像に対する認識結果（MUG）

図 3.4 は，完全未知のデータセット（RAVDESS データセット）の被験者に対して，それぞれの方法で認識を行った結果である．適用したデータは Anger 感情で歌を歌う様子を撮影したものであり，MUG データセットよりも表情が様々に変化するため，静止画像を対象としたモデルによる認識では認識結果がよりブレてしまっていることが分かる．一方，提案手法では，図 3.3 で示した結果と同様になめらかな認識を実現していることが確認できる．追加ネットワークは MUG に含まれる 52 名の被験者と比較的小規模なデータセットを利用して学習されているため，一般に未知の被験者に対する適用は難しい．しかし，静止画像を対象としたモデルが大規模データセットを利用して学習していることから，追加ネットワークは抽象化された特徴を入力として学習することができ，小規模データセットによる学習であってもロバスト性の高い学習が実現できていると言える．

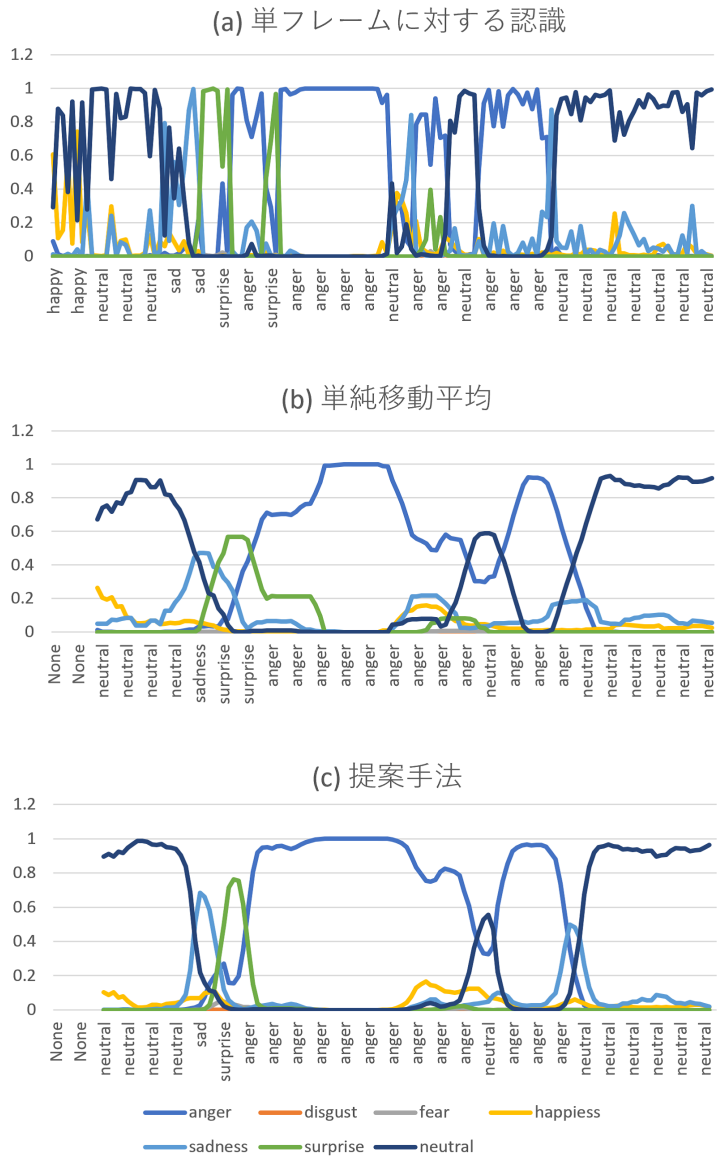

図 3.4: 動画像に対する認識結果（別データセット）

図 3.5 は図 3.4 の結果で利用した動画像データから，隣接フレームを抜き出したものであり，表 3.3 はそれぞれのフレームに対して静止画像を対象としたモデルと提案手法で認識を行った結果を表している．これらのフレームは，隣接フレームであることから非常に似通った表情であるため，これらの認識結果も近い値を取ることが望ましい．しかし，静止画像を対象としたモデルでは t と $t+1$ フレーム目の結果では結果が大きく異なっており，最大値を取る結果も Anger から Neutral に変化してしまっている．これは，まず静止画像を対象としたモデルは，1 枚の画像からの認識であり，前後のフレーム関係性を捉える機構がないことから，見かけ上微妙な表情変化であって抽出された特徴が変化した際に出力が大きく変化してしまうことが原因である．また，離散的なクラスが付与された静止画像を学習に利用していることから，離散的な出力（どこかが 1 に近づく出力）をするように学習が進んでいるため，ファジーな表情認識ができないことも原因の一つであると考えられる．一方，提案手法では複数フレームの関係性を捉える Attention 機構を導入したことで，隣接フレーム間の認識結果のずれが小さくなり，人間の感覚に近い認識を可能にしていることが伺える．

以上の結果から，提案手法では Attention 機構を有したネットワークを静止画像を対象としたモデルに導入することで，表情の時間方向の連続性をモデルが考慮することを可能にし，動画像に対する効果的な認識を実現していると言える．また，提案した学習方法は隣接フレーム間の認識結果のブレを抑制し，追加ネットワークは比較的少量のデータであっても学習可能であることが確認された．

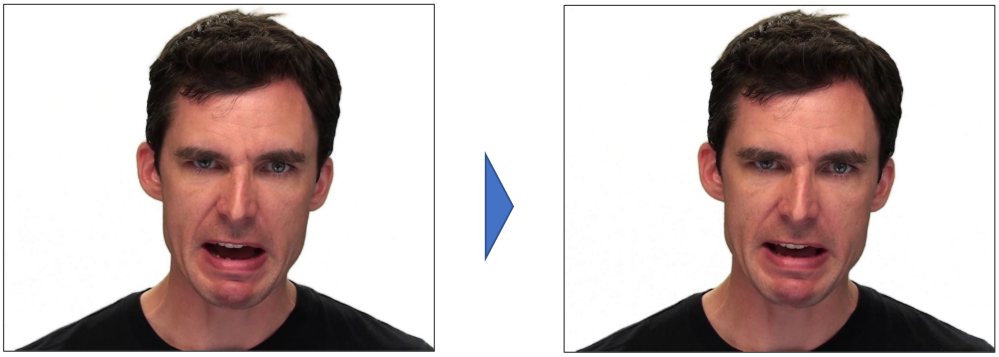

図 3.5: 隣接フレーム画像

表 3.3: 図 3.5 に対する認識結果

		anger	disgust	fear	happiness	sadness	surprise	neutral
単フレーム	t	0.718	0.007	0.009	0.110	0.013	0.013	0.129
	$t+1$	0.068	0.000	0.001	0.091	0.033	0.000	0.806
提案手法	t	0.825	0.002	0.003	0.108	0.029	0.010	0.023
	$t+1$	0.816	0.002	0.004	0.107	0.026	0.013	0.032

3.4 まとめ

本章では，静止画像を対象とした表情認識モデルに対して時間方向の連続性を考慮する機構を導入することで効果的に動画像へ適用する手法の提案を行い，実験でその有効性の検証を行った．

表情認識で利用するデータセットの多くは静止画像に対して離散的なラベルが付与されたものであり，そのようなデータを利用して学習した表情認識モデルは動画像に適用した際に，隣接フレーム間の微妙な表情変化によって認識結果が大きく変動し，認識として不自然な結果になってしまうといった課題があった．そこで，提案手法では静止画像を対象としたモデルに対して，Attention機構を有した追加ネットワークを導入し，時間方向の連続性をモデルが考慮できるようにすることで動画像に対して効果的な認識を実現した．また，この追加ネットワークの学習には教師情報を利用せず，隣接フレーム間の認識結果を近づけるように損失関数を設計することで，動画像に対するアノテーションの困難さを解決しつつ，静止画像を対象としたモデルで課題となっていた認識結果のブレを抑制することを可能にした．

第4章　表情ラベルを利用しない連続的な表情特徴の獲得

4.1　はじめに

第 3 章では，表情の時間方向の連続性を取り扱うことを目的として，静止画像を認識対象としたモデルに対し，時系列を考慮する追加ネットワークを導入し，動画像に対したモデルへの拡張を行った．この手法は，静止画像に対して離散的なラベルを付与して作成されたデータセットをモデルの学習に利用した際に起こる，表情の時間的な連続性の喪失に対して提案した解決手法であった．この章では，離散的なラベルと静止画像をモデルの学習に利用した際に起こる問題として新たに下記の 2 つを提起する．

- 表情の連続性を欠如した特徴の獲得
- 教師信号の曖昧性

離散的なラベルを教師とした教師あり学習では，一般的にモデル内で獲得される特徴をクラス間で引き離す方向で学習が行われる．これは表情認識についても同様であり，基本六感情の場合では，それぞれの感情は各々の特徴がまとまるように，そして感情クラスごとに特徴が離れるよう学習が行わる．しかし，表情は連続的に変化するものであるため，one-hot なラベルで表せない表情は多々存在する．つまり，少し怒っている表情や泣き笑いの表情など，表情は連続的に，そして複合的に表出する可能性があるが，それらは従来の one-hot なラベルを用いた教師あり学習ではそれらを扱うことができない．これにより画像生成のタスクでは中間的な表情の画像生成ができず，生成画像を利用したデータ拡張の際に可用性が低下したり，表情認識のタスクにおいてはファジーな表情認識の実現が困難になるなどの課題が生じる．

また，表情から受ける印象には人によって違いがあるため，教師信号であるはずのラベル自体が主観的なものとなり，曖昧性を持ったものになる可能性がある．これは，特に専門的な知識を必要しないアノテーション手法を利用し，素人によって構築された大規模データセットに起こりうる問題である．

これらの 2 つの問題は，表情に対してアノテーションを行うことで発生する課題であると言える．そこで，本章ではアノテーションによって与えられる表情に関する情報を利用することなく，表情の連続性を保持した特徴をモデル内部の潜在変数として獲得する手法の提案を行う．

4.2　Variational Autoencoder を用いた連続的な表情特徴の獲得

4.2.1　提案手法の概要

図 4.1 に提案モデルの概要を示す．提案モデルは VAE を拡張したものであり，IdentityEncoder と ExpressionEncoder の 2 つの Encoder と，共有の 1 つの Decoder から構成されている．IdentityEncoder

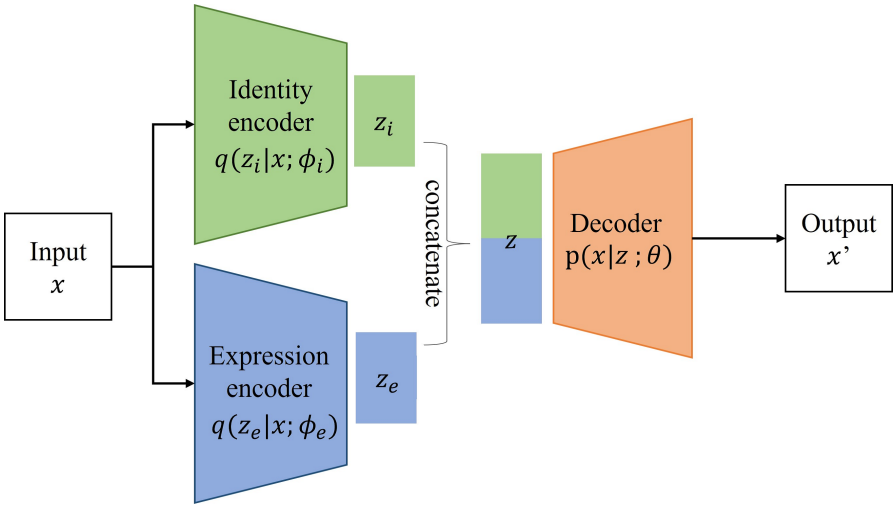

図 4.1: 提案モデル概要

と ExpressionEncoder は顔画像を入力として，それぞれの潜在変数 z_i・z_e に対して被験者特徴・表情特徴を分けた状態で埋め込むことを目的とし，Decoder はこれら 2 つの潜在変数から顔画像の生成を行う．このとき，ExpressionEncoder の潜在変数 z_e は被験者特徴から切り離された，純粋な表情特徴となることを期待する．

提案手法では，それぞれの Encoder にこれらの機能を具備するために「被験者特徴の獲得」「表情特徴の獲得」それぞれを目的とした，2 段階の学習を行う．このとき，表情に関する情報は学習には利用されず，顔画像がどの被験者の画像であるかという情報である被験者 ID のみを付加情報として利用する．

4.2.2 被験者特徴の獲得

被験者特徴の獲得を目的とした学習の概要を図 4.2 に，具体的な学習アルゴリズムを 1 に示す．このステージでは，被験者 ID を利用し，IdentityEncoder（$q(\boldsymbol{z_i}|x;\phi_i)$）と Decoder（$p(\boldsymbol{x}|\boldsymbol{z};\theta)$）の学習が行われる．一般的な VAE では式 2.2 で示したように，再構築誤差は入力画像 x とモデル出力 x' の間でとられるのに対し，ここでは式 4.1 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} 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が行われる．

$$\mathcal{L}_i = \alpha_{i1} D_{KL}(q(\boldsymbol{z_i}|x;\phi_i)||p(\boldsymbol{z})) - \alpha_{i2} E_{q(\boldsymbol{z}|x;\phi_i)}[\log p_\theta(x_{rand_{id}}|\boldsymbol{z})] \tag{4.1}$$

$\boldsymbol{z}$ は z_i と z_e を concatinate したベクトルを表し，id は被験者 ID を表している．α_{i1}，α_{i2} はそれぞれの項に対する重みパラメータである．またこのステージでは，ExpressionEncoder の学習は行われないため，z_e には，z_e と同じ次元数のゼロベクトルを代入する．z_e にゼロベクトルを代入する理由は，VAE は潜在変数の事前分布として正規分布 $\mathcal{N}(0, I)$ を仮定していることから，潜在変数において最も平均的な状態は 0 であると言え，また，このステージの学習の目的は被験者の特徴を獲得する事であることから，被験者特徴を獲得時の表情を最も平均的な状態として取り扱うためである．式 4.1 で示した損失関数では，同じ被験者の様々表情の間で再構築誤差を計算することになるため，モデルは入力画像の表情に影響されない特徴を潜在変数に埋め込むように学習が行われる．

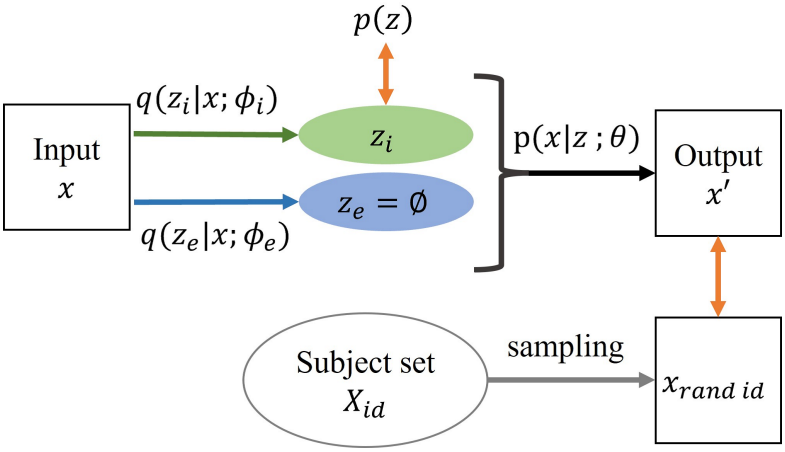

図 4.2: 被験者特徴獲得を目的とした学習

また，モデルは同じ被験者のすべての画像データの組み合わせに対して再構築誤差を小さくする必要があることから，同じ被験者の画像集合 X_{id} において最も尤度の高い画像を出力するようになる．このとき，IdentityEncoder の潜在変数 z_i には，表情の影響を受けない被験者の最も尤度の高い画像を出力するための特徴が獲得されていることが期待され，この特徴は被験者特徴であると考えることができる．

4.2.3 表情特徴の獲得

被験者特徴の獲得を目的とした学習は被験者特徴の獲得を目的とした学習の後に連続して行われ，モデルパラメータは引き継がれた状態で学習を始める．図**??**に表情特徴獲得の学習の概要を，学習アルゴリズムを 2 に示す．このステージでは，IdentityEncoder のパラメータ ϕ_i は学習済みパラメータとして固定され，ExpressionEncoder（$q(\boldsymbol{z_e}|\boldsymbol{x};\phi_e)$）と Decoder（$(p(\boldsymbol{x}|\boldsymbol{z};\theta)$）の学習のみ行われる．ここでは，前段の被験者特徴獲得で利用した損失関数とは異なり，式 4.2 で示すように入力画像 x とモデル出力 x' で再構築誤差をとる通常の VAE と同様の損失関数を利用する．

$$\mathcal{L}_e = \alpha_{e1} D_{KL}(q(\boldsymbol{z_e}|x;\phi_e)\|p(\boldsymbol{z})) - \alpha_{e2} E_{q(\boldsymbol{z}|\boldsymbol{x};\phi_e)}[\log p_\theta(x|\boldsymbol{z})] \tag{4.2}$$

前段の学習によって IdentityEncoder は被験者特徴を z_i に埋め込む機能を獲得しているため，モデルは入力画像の再構築を行うために，z_i に不足している情報を ExpressionEncoder の潜在変数 z_e に埋め込むように学習を行う．ここで，顔画像は被験者特徴と表情特徴によって構成されていると考えることができることから，ExpressionEncoder は z_e に表情特徴を埋め込む機能を獲得することになる．このとき獲得される表情特徴は被験者特徴から切り離された特徴であるため，仮に被験者が異なる場合であっても，似通った表情は近い値をとることが期待できる．

4.2.4 表情特徴の連続性

4.2.2，4.2.3 で説明を行った 2 段階の学習により，顔画像から被験者特徴と表情特徴を切り離した状態で獲得することが可能となった．ここでは，ExpressionEncoder の潜在変数 z_e として獲得された表情特徴が連続的な表情特徴になることについて説明を行う．

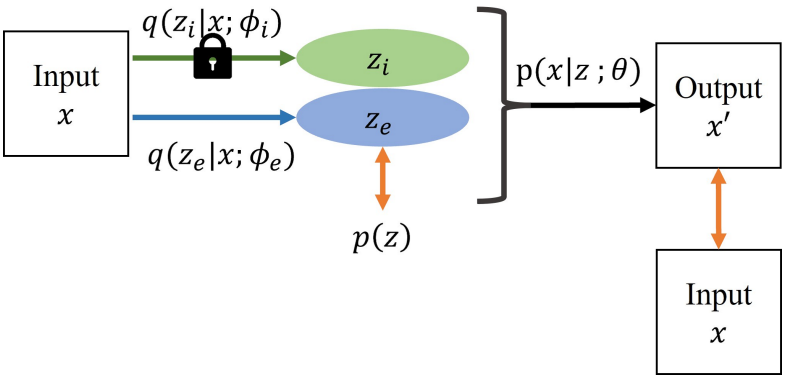

図 4.3: 表情特徴獲得を目的とした学習

Algorithm 1 Training algorithm : 被験者特徴の獲得

Require: Identity encoder : $q(z_i|\boldsymbol{x};\phi_i^0)$
Decoder : $p(\boldsymbol{x}|z_i,z_e;\theta^0)$
Image set : X

for $t = 0, ..., Titerations$ **do**
  Get N mini-batch samples from X
  for $x \in N$ **do**
    Get $x_{rand_{id}}$ samples from X_{id}
    Sample $z_i \sim q(z_i|\boldsymbol{x};\phi_i)$
    $z_e \leftarrow \emptyset$
    Decode z_i, z_e into $p(\boldsymbol{x}_{rand_{id}}|z_i,z_e;\theta)$
  end for
  Compute the loss using (4.1)
  Update $\phi_i^{t+1}, \theta^{t+1} \leftarrow \phi_i^t, \theta^t$ by gradient descent algorithm
end for

Algorithm 2 Training algorithm : 表情特徴の獲得.

Require: Expression encoder : $q(z_e|\boldsymbol{x};\phi_e^0)$
Identity encoder : $q(z_i|\boldsymbol{x};\phi_i)$
Decoder : $p(\boldsymbol{x}|z_i,z_e;\theta)$
Image set : X

for $t = 1, ..., Titerations$ **do**
  Get N mini-batch samples from X
  for $x \in N$ **do**
    Sample $z_i \sim q(z_i|\boldsymbol{x};\phi_i)$
    Sample $z_e \sim q(z_e|\boldsymbol{x};\phi_e)$
    Decode z_i, z_e into $p(\boldsymbol{x}|z_i,z_e;\theta)$
  end for
  Compute the loss using (4.2)
  Update $\phi_e^{t+1}, \theta^{t+1} \leftarrow \phi_e^t, \theta^t$ by gradient descent algorithm
end for

提案手法は VAE を拡張した手法であることから，学習時に Reparametrization Trick を利用して潜在変数の決定を行っている．そのため，潜在空間上で近い点に位置する特徴表現は似通った表現になりやすくなっている．その結果，ExpressionEncoder の潜在空間上では似通った表情は近い点にプロットされることが期待され，動画像などを利用してモデルの学習を行った際には，類似する表情特徴が潜在空間上で徐々に移動する形となる．逆に言えば，ExpressionEncoder の潜在空間上で近い点に位置する点は，似通った表情を表した特徴であると言える．この点から，z_e は連続的な表情特徴として捉えることができる．

4.3 表情特徴評価実験

実験の目的として以下の 2 点を設定する．

- 提案手法により顔画像に対して被験者特徴・表情特徴の切り分けが行われているかの検証
- ExpressionEncoder の潜在変数として連続的な表情特徴の獲得が行えているかの検証

ここでは，潜在空間の可視化と顔画像生成，表情認識の 3 つのタスクを行い，それぞれの検証を行う．

4.3.1 実験設定

データセット

本実験では以下の 3 つの動画像・画像列データセットを組み合わせて利用する．

MUG

The MUG Facial Expression Database [27] は 20 歳から 35 歳までの男性 51 人，女性 35 人、合計 86 人の被験者によって構成されたカラー動画像のデータセットである．表情は演技によって表出させたものであり，Neutral から基本六感情のうち 1 つの感情を表出したのち Neutral に戻る．本実験では一般に公開されている 52 人の被験者の動画像のみを利用した．

CK+

The Cohn-Kanade database [30] は 18 歳から 50 歳までの 123 人の被験者によって構成されたカラーまたはグレースケールの動画像データセットである．動画数は 593 本であり，そのうち 327 本にはアノテーションにより Ekman の基本 6 感情＋Neutral のラベルが付与されている．表情の表出は自発的に行われたものと演技によるものの両方が含まれており，Neutral から動画後半に向けて表情を表出させている．

RAVDESS

The Ryerson Audio-Visual Database of Emotional Speech and Song [31] は 21 歳から 33 歳までの男女 12 名ずつの被験者によって構成されたカラー動画像のデータセットである．俳優であるそれぞれの被験者が基本六感情に Neutral，calmness を加えた 8 つの感情を歌および表情によって表現している．

表 4.1: データセットの構成

	モデル学習用	実験結果提示用
被験者数	202 人（学習：195 人, テスト：7 人）	
データ数	30,351 枚 (学習：29,085 枚, テスト：1,266 枚)	2,037 枚 (学習：1,842 枚, テスト：195 枚)

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，モデル学習用と実験結果提示用の 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では，学習用の被験者の動画像から等間隔でフレームをサンプリングし作成する．これにより，表情が変化する途中の微妙な表情など様々な表情をデータセットに含めることが可能になり，モデルが表情の連続性を学習できるようにする．また，実験結果提示用のデータセットは，画像の表情と付与された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動画像からデータに割り当てられた表情が表出する箇所を自動的に切り出し作成する．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データセットとして利用し，RAVDESS は感情に紐づいた表情が表出する箇所を自動で判定できないため表情認識評価用のデータとしては利用しないこととした．

最終的な実験用データセットの構成を表 4.1 に示す．また画像は，学習済みの顔検出モデル［32］を使用して顔領域をトリミングしたのち，$64 \times 64 \times 3$ にリサイズされモデルに入力される．

モデル構造と学習設定

本実験で利用する Encoder と Decoder の構造を表 4.2，表 4.3 に示す．IdentityEncoder と ExpressionEncoder は同じ構造とし，潜在変数の次元数は 64 に設定した．モデルのパラメータは He らが提案した手法［29］を用いて初期化を行い，パラメータの更新には Adam［33］（$\beta_1 = 0.9, \beta_2 = 0.999, \sigma = 1.0 \times 10^{-8}, lr = 0.0005$）を用いた．学習のエポック数は被験者特徴の獲得・表情特徴の獲得それぞれのステージで 100epoch に設定し，学習率はエポックごとに 0.95 を乗算することで徐々に減衰させた．入力データの水増しとして，減色処理・鏡面反転・ガンマ補正をランダムに実施し，被験者特徴獲得の学習ステージでは，入力画像 x に実施した水増しと同様の処理がランダムサンプリング画像 $x_{rand_{id}}$ に施されるようにした．

表 4.2: IdentitiyEncoder , ExpressionEncoder の構造

Type	Kernel	Stride	Padding	Output
Image data	-	-	-	$3 \times 64 \times 64$
conv1_1	3×3	2	1	$32 \times 32 \times 32$
conv1_2	3×3	1	1	$32 \times 32 \times 32$
conv2_1	3×3	2	1	$64 \times 16 \times 16$
conv2_2	3×3	1	1	$64 \times 16 \times 16$
conv3_1	3×3	2	1	$128 \times 8 \times 8$
conv3_2	3×3	1	1	$128 \times 8 \times 8$
conv4_1	3×3	2	1	$256 \times 4 \times 4$
conv4_2	3×3	1	1	$256 \times 4 \times 4$
global_average_pooling	-	-	-	$256 \times 1 \times 1$
fc_μ	-	-	-	64
fc_σ	-	-	-	64

表 4.3: Decoder の構造

Type	Kernel	Stride	Padding	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	$16 \times 64 \times 64$
conv5	3×3	1	1	$3 \times 64 \times 64$

4.3.2 潜在空間の可視化

ここでは，IdentityEncoder と ExpressionEncoder の潜在空間を可視化することで，潜在空間にどのような表現が獲得されているのか確認を行い，被験者特徴と表情特徴が分離して獲得されているかの検証を行う．それぞれの Encoder の潜在変数の次元数は 64 であるため，直接可視化を行うことはできない．そこで，次元圧縮手法としてよく知られている t-distributed Stochastic Neighbor Embedding（t-SNE）[34]を用いて，潜在変数を低次元に圧縮することで，潜在空間の可視化を行う．

図 4.4 は IdentityEncoder の潜在空間の可視化したものである．ここでは，見やすさのため，202 人の被験者の中からランダムに 15 人（学習 : 13 人，テスト : 2 人）を選択し結果を表している．また，各点の色はそれぞれ被験者の ID ごとに割り当てられており，丸い点は学習用の被験者，四角い点はテスト用の被験者である．この結果から，IdentityEncoder の潜在空間上ではそれぞれの被験者ごとに特徴がまとまっていることが確認できる．これは，IdentityEncoder が表情が様々に変化する画像から被験者の特徴を抽出し，同じ被験者については表情が異なる場合であっても近い値の特徴として出力を行っているためである．この点から，IdentityEncoder は表情に影響されない被験者特徴を潜在変数として獲得しているといえる．

図 4.5 は ExpressionEncoder の潜在空間を可視化した結果であり，それぞれの点の色はデータに付与されたラベル情報と対応している．この結果から，ExpressionEncoder の潜在空間において，同じラベル付けられたデータはまとまっていることが分かる．今回利用したデータは，実験環境で撮影されたものであるため，異なるデータセットであっても，同じラベルがつけられたデータは似通った表情になっている．つまり，同じラベルがつけられたデータが潜在空間上でまとまるということは，ExpressionEncoder は被験者に影響されない表情特徴を抽出できていることを表している．また，緑色の縁で囲った fear のラベルが付けられたデータは，周囲のデータが Disgust であるため，一見，表情特徴の抽出がうまく行えていないように見える．しかし，実際のデータを確認すると，表情は Disgust の表情と類似していることから表情特徴の獲得は正常に行えていると言える．このようなラベルと表情が一致していないデータは，一般的な教師あり学習で学習を行った際に，意図しない学習を引き起こす原因となるが，提案手法では影響を受けずに表情特徴の獲得を実現できている．

4.3.3 顔画像の生成

この実験では，提案手法を用いて以下 3 つの顔画像の生成実験を行うことで，それぞれの Encoder の潜在変数 z_i，z_e の表現力や，2 段階の学習によって特徴の分離が目的通り実施できているかの検証を行う．

1. $z_e = \emptyset$，$z_i \sim \mathcal{N}(0, I)$ とした場合
2. $z_i = \boldsymbol{C}$（C は定数），$z_e \sim \mathcal{N}(0, I)$ とした場合
3. 被験者間で z_i，z_e を入れ替えた場合

まず 1 つ目の実験の目的は，表情特徴を固定しつつ被験者特徴のみをサンプリングすることで，獲得された被験者の多様性を確認するとともに，最も尤度の高い表情がどのような表情であるか検証することである．2 つ目の実験では 1 つ目の実験とは逆に，被験者特徴を固定しつつ表情特徴のみをサンプリングすることで，獲得された表情の多様性を確認することを目的としている．最後に 3 つ目の実験では，被験者間で表情特徴の入れ替えを行うことで被験者特徴と表情特徴の切り分けが

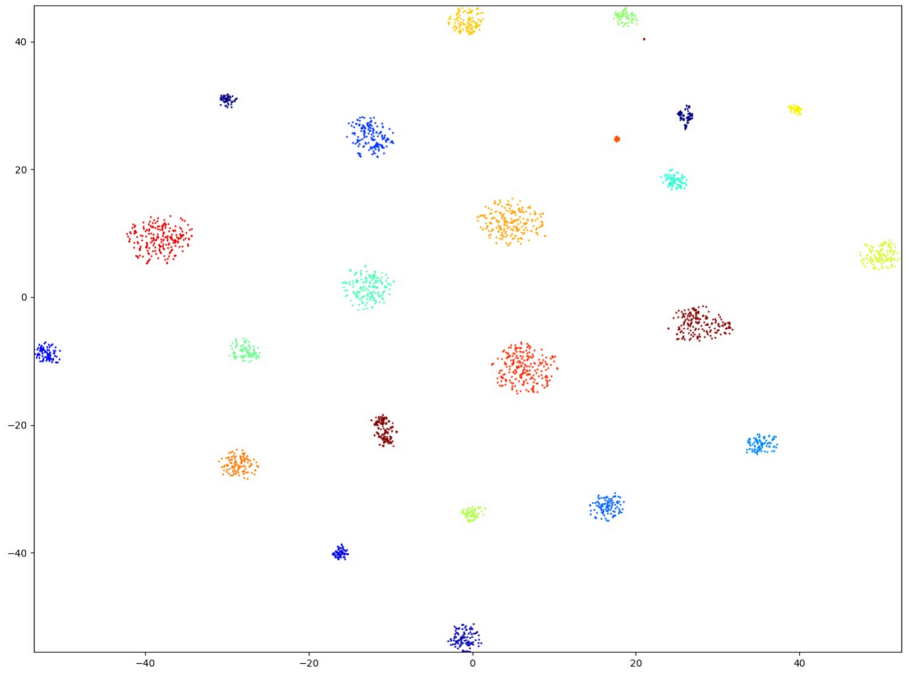

図 4.4: IdentityEncoder の潜在空間の可視化結果

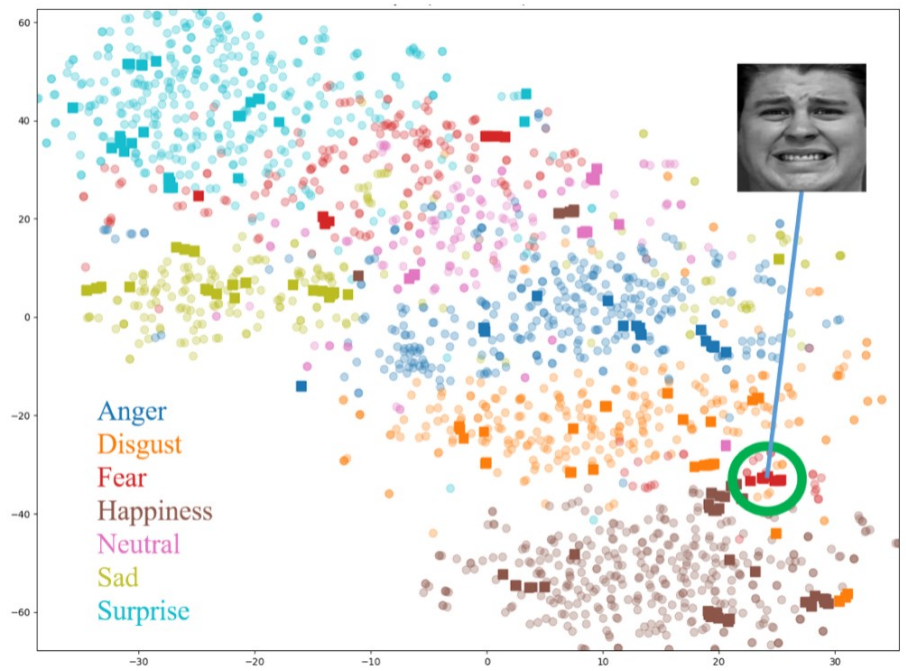

図 4.5: ExpressionEncoder の潜在空間の可視化結果

行えているかの検証が目的である．被験者特徴と表情特徴の分離が良好に行えている場合では，異なる被験者で特徴の入れ替えを行った場合でも，画像が崩れることなく被験者情報と表情特徴を維持した画像生成が可能となるはずである．

図 4.6 に 「$z_e = \emptyset$，$z_i \sim \mathcal{N}(0, I)$」とした場合の生成結果を示す．生成結果から提案モデルでは，男性や女性，髭の有無など様々な属性を持つ画像を生成できていることが伺える．この点から，IdentityEncoder の潜在空間には様々な属性を持つ被験者の特徴が獲得されていると言える．また，生成画像の表情に注目すると，すべての画像において Neutral に近い表情になっていることが分かる．これは，被験者特徴獲得の学習ステージにおいて，$z_e = \emptyset$ で学習を行っていたため，データの分布からモデル内部で被験者の最も標準的な状態として Neutral に近い表情が自然に獲得されたことを表している．一方で，生成結果として不自然な点も確認できる．まず，生成画像が崩れてしまっている点については，今回の実験で学習に利用した被験者が 197 人と少数であったため，被験者特徴の多様性を十分に獲得しきれなかったことが原因と考えられる．また，生成画像の色がカラーとグレーが混じったものになってしまっている点については，学習にカラー・グレースケール両方の画像を利用したため，z_i のサンプリングを行った際にその中間地点付近でサンプリングが行われたことが理由として挙げられる．

「$z_i = \boldsymbol{C}$（C は定数），$z_e \sim \mathcal{N}(0, I)$」とした場合の生成結果を図 4.7 に示す．$\boldsymbol{C}$ には，学習に利用した被験者の中からランダムに選択した画像の z_i を利用した．生成された画像には，喜怒哀楽など様々な表情が現れており，大きな表情だけではなく変化途中のような微妙な表情も確認できる．これは，学習に表情が連続的に変化する動画からサンプリングされた画像を利用していることから，モデルが表情特徴として様々なパターンの表情を獲得できたためと言える．

最後に,「被験者間で表情の入れ替えを行った」場合の生成結果を図 4.8・図 4.9 に示す．最左列と最上段は入力画像を表しており，縦方向は被験者特徴を固定，横方向は表情特徴を固定した結果である．まず図 4.8 に示した学習画像に対する結果に注目すると，被験者情報を維持したまま表情入れ替えを行うことができていることが確認できる．この点から，モデルは画像から被験者情報と表情特徴を分離した特徴表現として獲得できていることが分かる．次に図 4.9 に示したテスト画像に注目すると，被験者特徴が若干変化してしまっていることが確認できる．これは，1 つ目の実験結果で生成画像が崩れてしまったのと同様に，学習に利用した被験者の不足から，被験者特徴のパターンを十分に学習できなかったことが原因であると考えられる．しかし，表情についてはテスト画像についても良好に再現ができているため，ExpressionEncoder は被験者が未知になった場合であっても，顔画像から表情の特徴のみを抽出できていることが分かる．

以上の結果から，提案手法によって被験者特徴と表情特徴を分離した状態で獲得できていることが確認された．また，うまく生成が行えなかった画像について考察を行う中で，データセットの拡充，特に被験者数を増やすことで，提案手法が画像生成においてより良好な結果を示す可能性が示唆された．

図 4.6: 「$z_e = \emptyset$，$z_i \sim \mathcal{N}(0, I)$」とした場合の生成結果

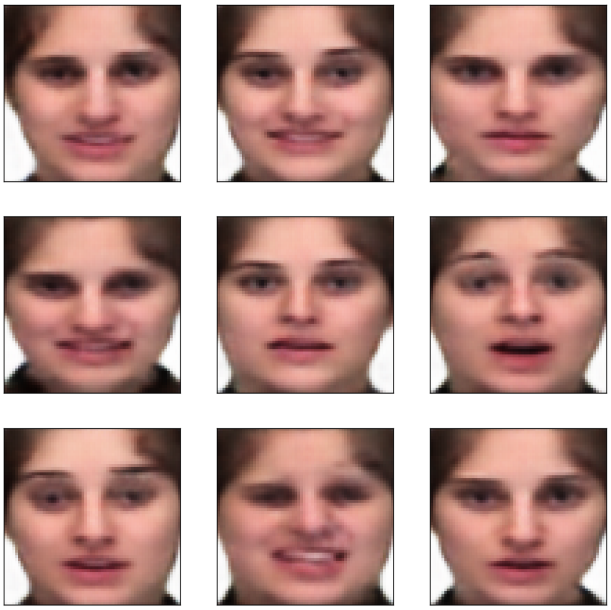

図 4.7: 「$z_i = \boldsymbol{C}$（C は定数），$z_e \sim \mathcal{N}(0, I)$」とした場合の生成結果

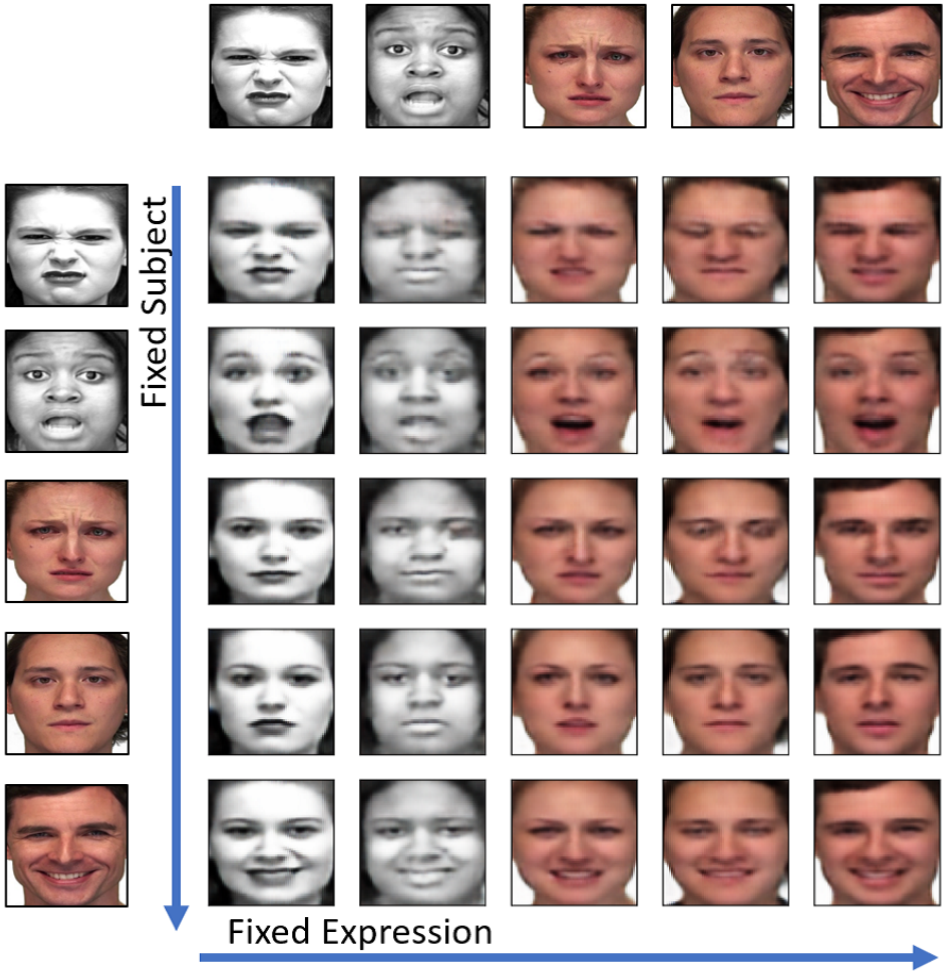

図 4.8: 学習データに対する表情入れ替え画像生成結果

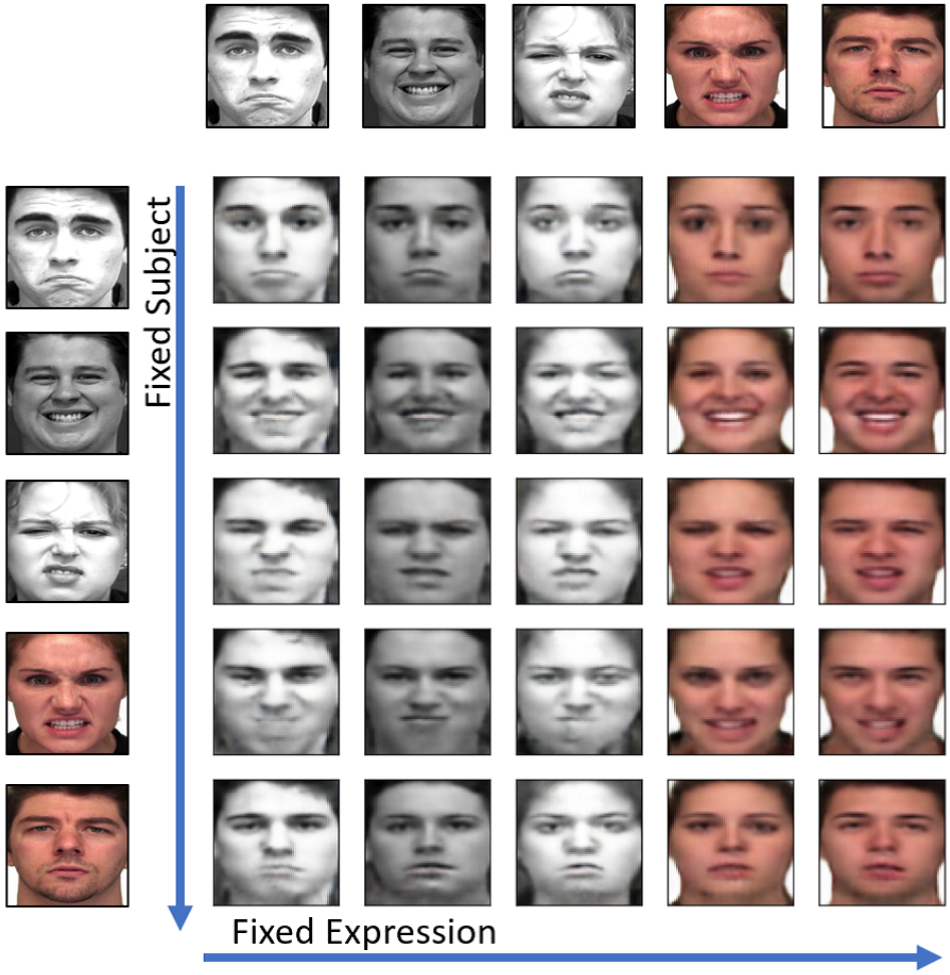

図 4.9: テストデータに対する表情入れ替え画像生成結果

4.3.4 潜在空間を利用した表情認識

4.3.3 の実験結果より，提案手法は顔画像から被験者特徴と表情特徴に分離した状態で，潜在変数として獲得できることが分かった．ExpressionEncoder の潜在変数 z_e は純粋な表情特徴であることから，表情認識に対して有効であると考えられる．また，提案手法は VAE を拡張した手法であることから，ExpressionEncoder の潜在空間上で近い点は似通った表情であり，提案手法によって良好に表情特徴が抽出できていれば，単純なユークリッド距離によって表情の類似度を測定することが可能であるといえる．そこでここでは，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によって表情認識を行うことで，獲得された表情特徴の有用性の検証を行う．今回の実験では，クラスタリング手法として k-means++ [35]を利用し，クラスタ数 $K = 9$ とした．実験設定としては，基本六感情+Neutral の 7 クラス分類問題であるが，一般的にクラスタ数はクラス数に対してゆとりを持たせるため，今回は実験的に $K = 9$ としている．クラスタに対するクラスの割り当てにはセントロイドから最も近い点の画像につけられたラベルを利用した．また，k-means++は初期のセントロイドの決定にランダム性が含まれており，それによりクラスタリングの結果が変わるため，本実験では認識精度を出す際にクラスタリングを 100 試行し，その平均を算出した．

表 4.4 に提案モデル（ExpressionEncoder）とシンプルな VAE のそれぞれの潜在空間を用いた場合の表情認識の結果を示す．シンプルな VAE は提案手法と同様の Encoder と Decoder の構造を用いて構成されたものである．この結果から，提案手法はシンプルな VAE と比べて高い認識精度を実現していることが分かる．これは，シンプルな VAE では潜在変数として獲得される特徴には表情特徴以外の特徴が含まれ，それらが複雑に絡み合った特徴となっているため，単純なユークリッド距離によって表情の類似度を測るのが難しいのに対して，提案手法では ExpressionEncoder が表情特徴のみを獲得しつつ，VAE の特性を利用して類似した表情を潜在空間上で近い点にプロットしているためである．また，提案手法ではテストの被験者に対しても精度の低下が抑えられている点から，表情特徴を被験者に依存したものではなく被験者間で共通したものとして獲得できていることが伺える

表 4.4: 潜在空間を用いた表情認識結果 (100 試行平均)

	提案モデル (Expression encoder)	Simple VAE
学習	56.33 ± 2.54%	38.08 ± 1.56%
テスト	49.95 ± 3.58%	22.07 ± 3.21%

図 4.10 は潜在空間においてどのようなクラスタが形成され，認識が行われたのかを次元圧縮手法によって示したものである．最上段の図 4.10(a) はデータに付与されたラベルごとに色付けがなされており，表情認識の正解を表している．中段の図 4.10(b) はクラスタの結果を表しており，クラスタごとに色が割り当てられている．最下段の図 4.10 は表情認識の結果を表している．つまり，最上段と最下段で色の分布が類似しているほど，表情認識が良好に行われたことを表している．まず，シンプルな VAE の結果に注目すると，(a) の結果では VAE の潜在空間では表情としてまとまりはほとんど形成されておらず，小さなまとまりをいくつも形成していることから特徴が被験者でまとまってしまっていることが考えられる．そのため，(a) と (c) の色の分布の一致率は非常に低い．一方提案手法では，潜在空間で表情ごとにまとまりを見せていることから，良好にクラスタの形成と，認識を行えている．また，提案手法において (a) と (c) の結果を比較すると，誤認識をし

ている箇所としては主に，ラベル（色）の境界線付近のみであり，色の分布は非常に似ていることから，表 4.4 で示した精度以上に獲得された表情特徴の有効性は高いことが示唆される．

以上の結果から，獲得された表情特徴の表情認識に対する有効性が示唆された．実験で行った表情認識は，クラスタリングによるものであり，モデルの学習にも表情に関するラベルは用いていないことから，教師なし学習による表情認識を実現しているといえる．また，今回は獲得された潜在空間をクラスとして区切り認識を行ったが，セントロイドからの距離によるファジーな表情認識などにも応用できることから，提案手法で獲得された表情特徴の有用性は非常に高いと言える．

図 4.10: クラスタリング及びクラス分類結果

4.3.5　追加実験：別手法による潜在空間の可視化

4.3.2 では，次元圧縮手法の t-SNE を利用し ExpressionEncoder の潜在空間を実施した．次元圧縮は様々な仮定をおき，本来であれば高次元で表現されるものを低次元に落とし込む方法であるため，実際に獲得された特徴と可視化によって確認された情報では乖離がある可能性がある．そこで，ここでは t-SNE とは異なる次元圧縮手法 (UMAP:Uniform Manifold Approximation and Projection［36］) を利用し，パラメータを変更した場合の可視化を行うことで，ExpressionEncoder の潜在空間に獲得された特徴のより深い理解を行う．

図 4.5 はそれぞれのパラメータ設定で UMAP によって次元圧縮を行った結果を表している．また，距離関数にはユークリッド距離を利用している．設定したパラメータの *n_neighbors* は多様体構造の局所近似で利用される隣接点の数を表し，値を小さくするとローカルな構造を、値を大きくするとグローバルな構造を確認することができる．また，*min_dist* は埋め込みを行う際に，点同士の圧縮の程度を表しており，値を小さくすることでローカルな構造に関して正確に最適化を行うことができ，値を大きくすると点同士を分散して最適化を行う．まず *n_neighbors*，*min_dist* ともに小さな値を設定した，ローカルな構造に対する可視化結果に注目すると色が局所的にまとまり，クラスタを形成していることが分かる．この点から，Encoder の潜在空間では表情クラスのクラスタが形成されていることが伺える．一方，グローバルな構造に注目すると，t-SNE で可視化した場合と同様に，色はまとまりながらも点が広がっていることが分かる．これは，Encoder の潜在空間が離散的なものではなく，連続的な構造を取っていることを表しており，連続的な表情の特徴の獲得が実現できていると考えらえる．

以上の結果から，Encoder の潜在空間では表情クラスは比較まとまった表情クラスタを形成しており，また獲得されている表情特徴については表情の連続性を保有したものであることが確認された．

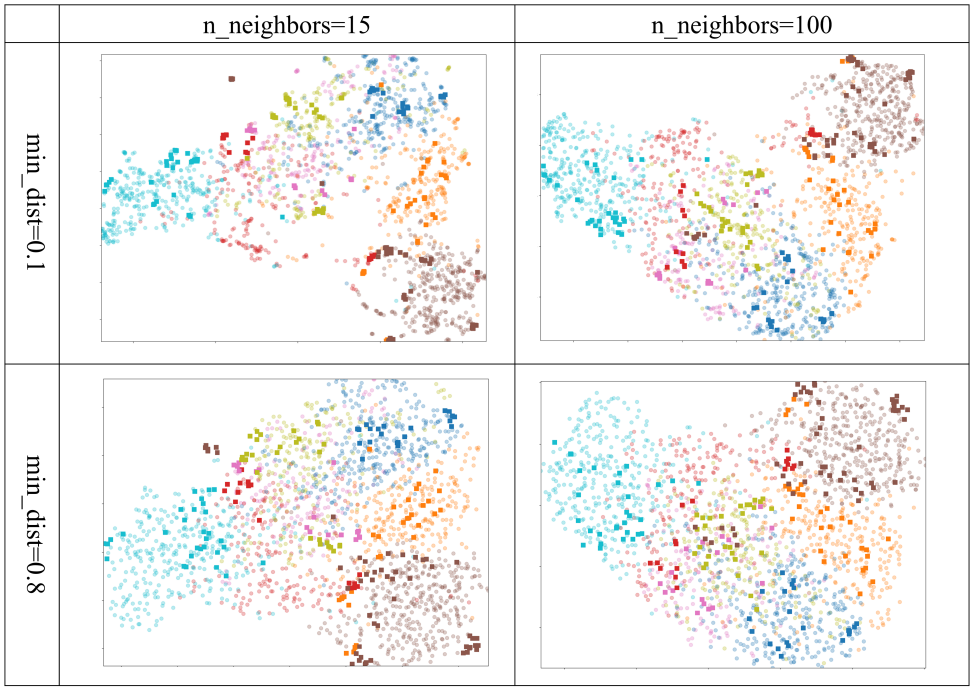

図 4.11: UMAP による ExpressionEncoder の潜在空間の可視化結果

4.4 まとめ

本章では，表情に関する情報を用いることなく連続的な表情特徴を獲得する手法の提案を行うとともに，実験によりその特徴の有用性を示した．

一般に機械学習において表情を扱う場合，離散的なラベルを利用したアノテーションが必要となるが，それによって表情という連続的な事象を離散的に扱うことになったり，教師信号としてのラベル自体が曖昧性を持ったものになるなど問題があった．そこで提案手法では，VAE の枠組みを拡張し，被験者情報を付加情報として学習に利用することで，モデルの潜在変数として，被験者特徴と表情特徴を分離した状態で獲得することを行った．また，実験では，潜在空間の可視化・顔画像生成・潜在空間を利用した表情認識を行うことで，提案手法のそれぞれのタスクに対する有効性を示すとともに，獲得された表情特徴の有用性の検証を行った．

第5章　詳細な要素を捉えた連続的な表情特徴の獲得

5.1　はじめに

第 4 章では，表情に対するアノテーションによって引き起こされる問題の解決を目的として，表情に関する情報を用いずに連続的な表情特徴を獲得する手法の提案を行い，有効性の検証を行った．しかし，第 4 章で提案した手法（以後”先行手法”と呼ぶ）では学習に乱数を利用する VAE の特性上，生成画像がぼやけてしまっていることから，顔のしわなどの表情を構築する上で重要な細かな要素が表情特徴して捉えられていないことが示唆されていた．そこで本章では，先行手法に対して改良を行うことで，より詳細な表情の要素をモデルの潜在変数として獲得する手法の提案を行う．以下，提案手法について詳しく述べて行く．

5.2　生成画像の鮮明化と詳細な表情特徴の獲得

本手法の目的は先行手法を改良し，より効果的な表情特徴を獲得することである．ここで効果的な表情特徴とは，しわなどの表情を構成する細かな要素を捉えながら，被験者間で共通な表情を表現する特徴をさす．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の見直しを行うことで，モデルの過学習を抑制しながら鮮明な顔画像の生成を行うことで，効果的な表情特徴の獲得を実現する．

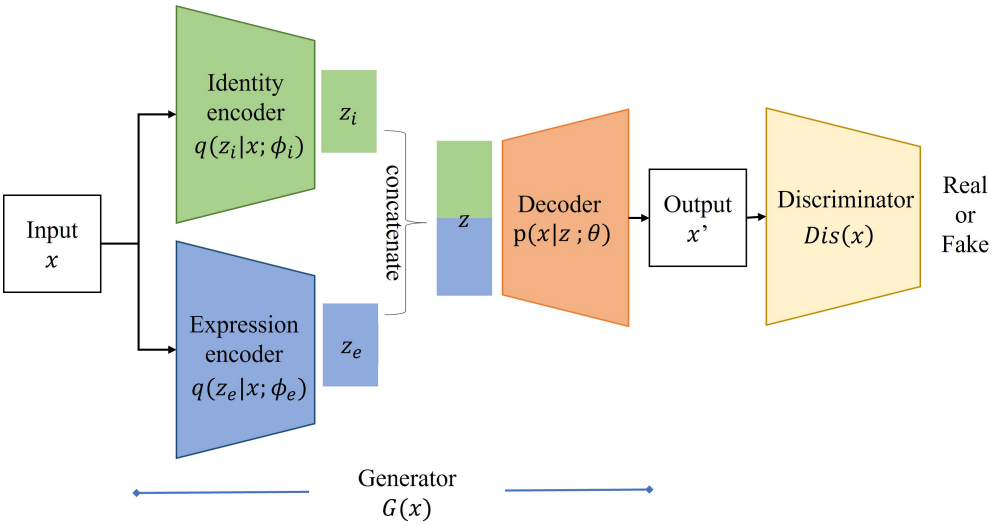

図 5.1: 提案モデル概要

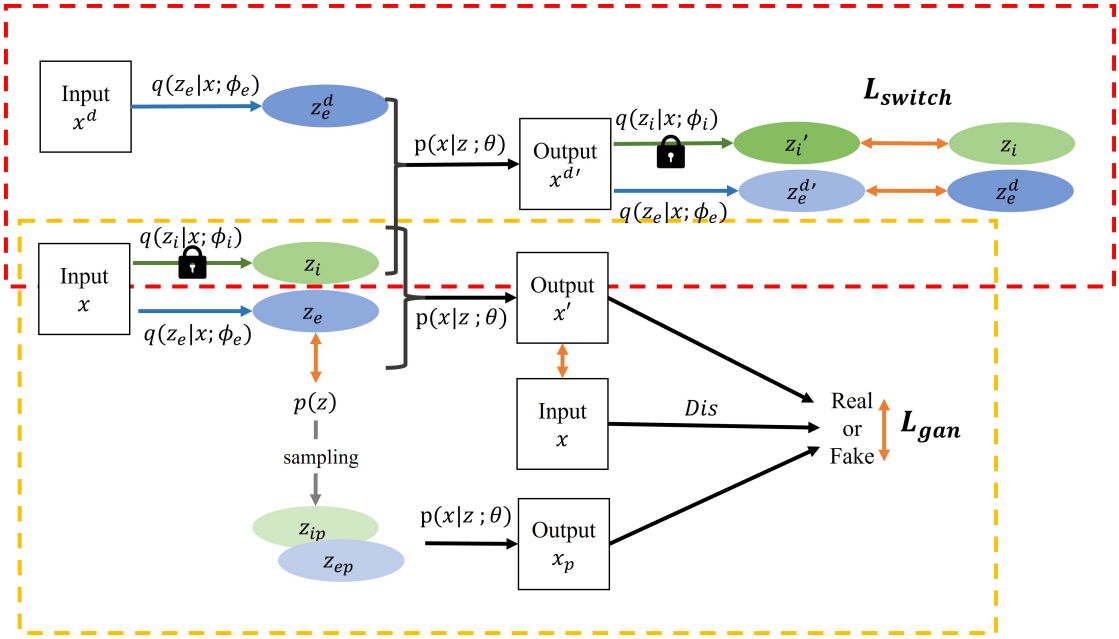

図 5.2: 表情特徴獲得を目的とした学習（改良）

5.2.1 損失関数の改良

ここでは，効果的な表情特徴の獲得を目的として導入した 2 つの損失関数について説明を行う．効果的な表情特徴の獲得には「生成画像の鮮明化」と「表情特徴が被験者に依存しないよう抑制する」ことが重要と考え，それぞれを目的とした損失関数を導入する．これらは，表情特徴の獲得に対して効果が期待される損失関数であることから，被験者特徴の獲得を目的とした学習ステージには利用せず，表情特徴の獲得を目的とした学習ステージのみに導入する．

Adversarial Loss

Adversarial loss は Generative adversarial networks（GAN）[23]で提案され，生成のタスクを行う際に多く用いられる，Generator と Discriminator の 2 種類のネットワークを利用した損失関数である．Generator は生成を担うネットワークであり，Discriminator は入力されたデータが本物のデータか Generator によって生成されたデータかを分類するネットワークである．Adversarial loss はこれらのネットワークが互いに影響を与え合うように設計されており，Generator がよりリアルなデータの生成を行うことを可能にする．つまり，Generator は Discriminator に分類が行えないように生成を行い，Discriminator は分類を正確に行えるように学習をおこなうため，結果として Generator の生成結果は本物のデータに近いものになることが期待される．

提案手法ではこの性質を利用し，生成される顔画像を鮮明化することで，より細かな表情の特徴を潜在変数に獲得する．図 5.2 の黄色の点線で囲われた箇所は，今回導入した Adversarial loss の概要を表している．提案モデルでは図 5.1 に示すように，2 つの Encoder と Decoder で構成された箇所を Generator（G）として扱い，新たに Discriminator（D）を導入する．

$$\mathcal{L}_{gan} = \min_{G} \max_{D} \{\log(D(x)) + \log(1 - D(x')) + \log(1 - D(x_p))\} \tag{5.1}$$

式 5.1 は提案手法で利用する Adversarial loss を表したものである．x は実画像（本物の顔画像），x' は Generator によって生成された画像，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潜在変数 z として Decoder に入力した際の出力画像である．サンプリングされた潜在変数による生成結果に対しても，Adversarial loss を算出することで，潜在空間の表現力をより高めることが可能となり，実画像の潜在空間上での射影点以外の値でもリアルな画像の生成が期待できる．

Switch Loss

図 5.2 の赤枠で囲われた部分は、今回導入を行った損失関数の一つである Switch loss の概要を表している．Switch loss は被験者特徴と表情特徴が絡み合った状態で潜在変数に獲得されることを抑制しつつ，被験者間で共通した表情特徴を獲得することを目的として導入する損失関数である．ここで被験者間で共通した表情特徴とは，被験者が異なっている場合で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．

前述した Adversarial loss のみを損失関数として表情特徴獲得の学習ステージに導入した場合，モデルの学習は画像の鮮明化を目的とした方向に進みやすくなる．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，ExpressionEncoder の潜在変数 z_e のみに依存して画像の再構築を行うようになることが予測される．その結果，z_e に表情以外の特徴が入り込み，表情特徴が被験者に固有のもの，もしくは被験者特徴と表情特徴が絡み合った特徴になることが考えられる．そこで，そのような Adversarial loss のデメリットを抑制し，被験者間で共通した表情特徴を獲得するために，式 5.2 に示すような Switch　loss を導入する．

$$\mathcal{L}_{switch} = MSE(z_i, z_i') + MSE(z_e^d, z_e^{d'}) \tag{5.2}$$

ここで，z_i はある画像から IdentityEncoder によって抽出された被験者特徴を表しており，z_e^d は z_i の抽出に利用した被験者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 に入力した際の出力 $x^{d's}$ をもう一度 IdentityEncoder と ExpressionEncoder に入力した際の潜在変数の値を表している．Switch loss は元画像から抽出された潜在変数と生成画像から抽出された潜在変数間で誤差を算出する関数であり，表情特徴 z_e が異なる被験者間で入れ替わった場合でも，Decoder の出力に入れ替える前の表情特徴を維持することを意図して設計された関数である．これにより，表情特徴を被験者間で共通したものとして抽出するよう学習を進めることができるようになる．また，Switch　loss では被験者特徴についても z_i と z_i' の間で誤差を算出しているため，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た生成を行うことが抑制される．これにより，獲得される特徴が被験者特徴と表情特徴として分離されたものになり，被験者特徴間で共通した表情特徴として ExpressionEncoder の潜在変数に獲得することが可能となる．

表情特徴獲得ステージの損失関数

提案手法では，4.2.3 で説明を行った損失関数 L_e に，新たに Adversarial loss と Switch loss を導入し，学習を行う．

$$\mathcal{L}_e' = \mathcal{L}_e + \beta\mathcal{L}_{gan} + \gamma\mathcal{L}_{switch} \qquad \beta, \gamma > 0 \tag{5.3}$$

β，γ はそれぞれ Adversarial loss と Switch loss の重みパラメータを表しており，それぞれの損失関数が学習に与える影響を制御するために利用される．

5.2.2　学習ステップの改良

IdentityEncoder の潜在変数 z_i と ExpressionEncoder の潜在変数 z_e は互いに異なる情報を保持し，互いの欠損する情報を補完しあうことで顔画像の生成を可能にしている．このことから，それぞれのエンコーダの学習は，互いに影響を与え合いながら行われることが望ましいと考えられる．しかし，先行手法で行われた学習では，図 5.3 に示すように，それぞれの学習ステージにおいて一方のエンコーダの学習のみ行われており，また学習は一巡であったために，第 1 ステージの被験者特徴獲得で学習が行われる IdentityEncoder は ExpressionEncoder の学習結果を加味して，自身の学習を行うことができていなかった．そこで提案手法では，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により，ExpressionEncoder の学習結果の影響を IdentityEncoder の学習に間接的に伝える．ここで「間接的」と表現するのは，被験者特徴獲得のステージでは潜在変数 z_e にはゼロベクトルが代入されており，ExpressionEncoder は直接学習には関わらないが，ExpressionEncoder の学習結果は表情特徴獲得ステージ後の Decoder のパラメータによって伝えられるためである．

またこの学習ステージの繰り返しは，5.2.1 で導入した Adversarial loss と組み合わせることで，被験者特徴と表情特徴を分離しつつより生成画像を鮮明化する効果も期待できる．Adversarial loss を利用して鮮明な画像を生成しようとすると，一般に学習エポック数を増やす必要が出てくる．しかし，先行手法のように学習ステージを一巡回すだけでこれを行おうとすると，後段に行われる表情特徴の獲得を目的とした学習ステージのエポック数を増やすことになるため，ExpressionEncoder と Decoder のみに依存した学習が進められる．その結果，z_i の情報をほとんど利用せず z_e の情報のみで生成が行われるため，被験者特徴と表情特徴が絡み合った特徴になることが考えられる．一方，学習ステージを繰り返し行う方法では，表情特徴獲得の学習ステージのエポック数を増すことで生成画像の鮮明化を行いつつ，IdentityEncoder の学習も繰り返しごとに行われることから，一方のエンコーダに依存した学習が起こりにくくなり，特徴を分離した状態で獲得することが可能になる．

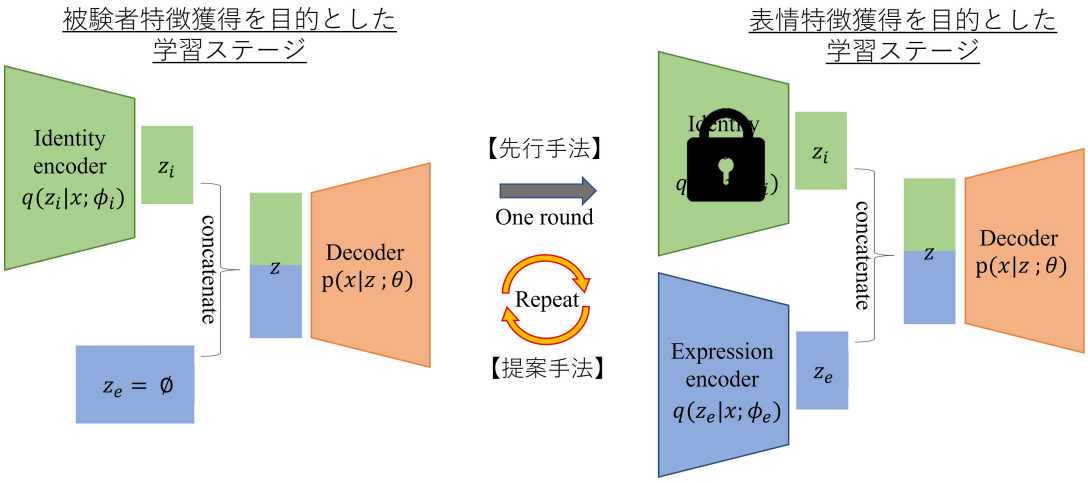

図 5.3: 先行手法/提案手法における学習ステージ

5.3 表情特徴評価実験

ここでは，提案手法により効果的な表情特徴の獲得が実現されることを検証することを目的として，第 4 章にならい，表情認識と画像生成の 2 種類のタスクを実施する．表情認識のタスクでは，ExpressionEncoder の潜在変数を利用して，単純な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モノとして獲得されているか評価を行う．また，顔画像生成のタスクでは，表情の入れ替え（Swapping）を行うことで被験者特徴と表情特徴の分離度を評価し，ある表情からまたある表情の間の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の評価を行う．

5.3.1 実験設定

データセット

本実験で利用するデータセットは，第 4 章の実験で利用したものと同様の表 4.1 で示した構成のデータセットを利用する．ただし，今回は検証の中で入力画像サイズを変更した場合を行うため，顔検出モデルによって顔領域のトリミングを行った後，検証内容に応じて $64 \times 64 \times 3$ または $128 \times 128 \times 3$ にリサイズしモデルに入力される．

モデル構造と学習設定

本実験で利用する Encoder と Decoder，Discriminator の構造をそれぞれ表 5.1，表 5.2，表 5.3 に示す．表中に括弧で示した箇所は，入力サイズが $128 \times 128 \times 3$ の場合に利用されるレイヤーと出力サイズを表しており，入力サイズが $64 \times 64 \times 3$ の場合は，第 4 章で行った実験のモデルと同様の構造を利用する．また，学習設定についても前章と合わせて実験を行った．モデルの学習パラメータは，He らが提案した手法により初期化され，パラメータの更新には Adam（$\beta_1 = 0.9, \beta_2 = 0.999, \sigma = 1.0 \times 10^{-8}, lr = 0.0005$）を利用した．学習エポックは学習ステージを一巡のみ回す場合は各学習ステージ 100epoch とし，学習ステージを繰り返し回す場合には各学習ステージで 50epoch とした．また，画像の水増し関数として，入力画像に対して減色処理・鏡面反転・ガンマ補正をランダムに適用し，被験者特徴獲得の学習ステージでは，入力画像に適用した水増し処理と同様の処理がランダムサンプリング画像 $x_{rand_{id}}$ に施されるように設定した．目的関数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．

表 5.1: Encoder の構造

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 5.2: Decoder の構造

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	$3 \times 64 \times 64$ ($3 \times 128 \times 128$)

表 5.3: Discriminator の構造

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_1	-	-	-	1

5.3.2 潜在空間を利用した表情認識

この実験では，潜在空間を利用したユークリッド距離によるクラスタリングを行うことで，獲得された表情特徴の評価を行う．提案手法によって ExpressionEncoder の潜在変数として被験者に共通した表情特徴の獲得が行えている場合，似通った表情は ExpressionEncoder 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ても小さくなるため，表情認識は良好に行えるようになると言える．そこで，今回はユークリッド距離によるクラスタリング手法として k-means++を用い，様々な手法の潜在空間を利用して表情認識を行い，結果を比較することで提案手法の表情特徴の評価を行う．k-means++のクラスタ数を決定するパラメータは $k = 9$ とし，クラスタに対するクラスの割り当てには各クラスタのセントロイドに最も近い点の画像に付与されたラベルを利用した．また，今回の精度比較は 100 試行平均の値を利用して実施した．比較手法として，まず第 4 章の表情認識実験で利用した，シンプル VAE と先行手法を設定する．シンプル VAE は被験者情報や特殊な学習方法を利用しない場合のベースラインの精度を知ることを目的としており，先行手法は今回実施した改良が精度に対してどのような影響を与えるかを検証することを目的としている．また，被験者情報を利用した場合の既存手法による精度を確認するため，被験者 ID をラベル情報として用いた場合の C-VAE [37]の潜在空間を利用した場合と，シンプル VAE の潜在変数に対して式 5.4 に示した処理を施した場合の精度を確認した．式 5.4 は被験者ごとの潜在変数の平均を取ったものと被験者特徴として扱い，それをそれぞれの潜在変数の値から減算することで，被験者特徴を潜在変数として獲得された特徴から除く処理を施したものである．

$$
\begin{aligned}
f_{sub_{ij}} &= x_{ij} - \frac{1}{n}\sum_{k=1}^{n} x_{ik} \\
&i : subject\ identity\ number \\
&x_{ij} : the\ jth\ image\ in\ the\ subject\ set\ X_i
\end{aligned}
\tag{5.4}
$$

表 5.4 は各手法の潜在空間を利用した場合の表情認識の結果を表している．ここで，train はモデルの学習に利用した被験者，test は学習に利用していない被験者の結果であり，どちらも教師なしで表情認識を行った結果を表している．また，⑦・⑧の手法では被験者 ID をモデルの入力や特徴量の算出に利用しているが，test は被験者情報が未知の設定であることから手法の適用を行うことができないため，test の結果が空欄になっている．

まず最終的な精度に着目すると，提案手法④と先行手法①，比較手法⑥⑦⑧の結果を比較により提案手法が最も高精度に表情認識を行えていることが分かる．このことから，提案手法が被験者情報を利用した学習において最も効果的な表情特徴の獲得を実現できていることが分かる．次に，提案手法で導入した損失関数の効果について先行手法の結果と比較しながら確認して行く．②の Adversarial　loss のみを導入した場合では，先行手法よりも精度が低下してしまっていることが分かる．これは，Adversarial loss のみを導入した場合では，表情特徴獲得のステージの学習の影響が大きくなることで，IdentityEncoder の学習結果を無視し，ExpressionEncoder と Decoder のみで画像の再構築を行うように学習が行われたため，被験者特徴と表情特徴が絡み合った状態で獲得されてしまったことが原因であると考えられる．一方，③の Switch loss を導入した場合では，Switch loss の導入の意図通りに Adversarial loss のデメリットを解決したことで精度向上に転じていることが確認できる．さらに，提案手法では Adversarial loss 導入により，詳細な表情の特徴を捉えることができるようになっているため，入力画像のサイズを 64 から 128 に大きくすることで，より細かな表情の特徴を学習時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，④・⑤の結果を比較することで分かる．

学習ステージの繰り返しの効果に着目すると，④の結果が③の結果を上回っていることから，学習ステージの繰り返しによってそれぞれの Encoder が互いに影響を与え合いながら学習が進み，より効果的な表情特徴の獲得を実現できていることが伺える．また，図 5.4 は入力サイズを 128 とした時の学習ステージの繰り返しによる表情認識精度の変化を表したグラフである．縦軸は精度，横軸は学習ステージの繰り返し回数を表している．この結果から，学習ステージの繰り返しにより徐々に精度が上がっていき，1 巡による学習よりも高い精度を実現していることが分かる．これは，学習ステージを交互に行うことで，各 Encoder が繰り返しごとに他方の Encoder の学習結果を加味した学習を行ったことで，損失関数と学習方法の効果が高まり，より明確に被験者特徴と表情特徴を分離する方向で学習が進んだためであると考えられる．

以上の結果より，提案手法で行った 2 種類の損失関数の導入と学習ステップの改良により，被験者特徴と表情特徴が分離され，被験者間で共通した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．

表 5.4: 表情認識結果

	input size	train	test
①conventional method［38］	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%	-

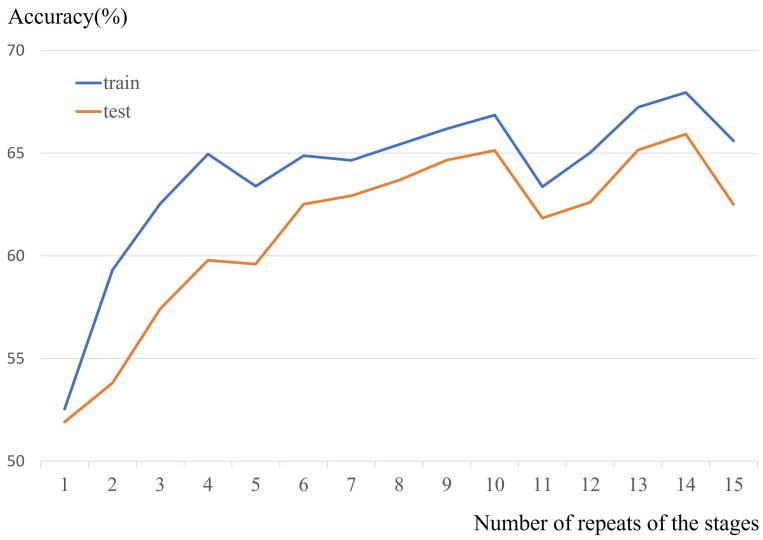

図 5.4: 学習ステージの繰り返しによる認識精度の変化

5.3.3 顔画像の生成

この実験では顔画像生成として，異なる被験者間での表情の入れ替え（Swapping）と，異なる 2 種類の表情の補間（Interpolation）を行うことで，潜在変数として獲得された被験者特徴と表情特徴の分離度と，表情特徴の連続性の評価を行う．

表情の入れ替えは，ExpressionEncoder の潜在変数 z_e を異なる被験者と交換した特徴を Decoder に入力し，画像生成を行うことで実施する．獲得された特徴が被験者特徴と表情特徴で分離された特徴であれば，表情特徴を入れ替えた場合であっても被験者や表情の情報が変化することなく画像生成を行うことができるはずである．

図 5.5，5.6 はそれぞれ，学習データとテストデータに対して表情入れ替えを行った結果を表している．最も左の画像は表情特徴を抽出する画像，最も上段の画像は被験者特徴を抽出する画像を表しており，それぞれの列は先行手法・提案手法（繰り返し学習なし）・提案手法（繰り返し学習あり）の各手法による生成結果を表している．また，括弧内の数字は入力画像の大きさである．まず，図 5.5 の学習データに注目すると，先行手法では生成画像がぼやけていたり，埋め込みを行った際に生成画像が崩れてしまっていることが分かる．一方，提案手法では生成画像が鮮明になっており，しわなど細かな表情の要素まで生成画像に再現することが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の比較を行うと，生成画像の崩れが小さくなるとともに，眉間や口周りのしわなどより鮮明に再現できていることが分かる．これらの結果から，提案手法で行った改良により，生成画像の鮮明化を実現したのみではなく，被験者特徴と表情特徴をより分離し，より細かな表情の要素を捉えた表情特徴を獲得できていることが確認された．図 5.6 のテストデータに対する結果に注目すると，すべての手法において被験者の特徴が変化し別人の画像生成を行ってしまっていることが分かる．これは，本実験で学習に利用した被験者数が 195 名程度であったために，被験者の特徴を十分に網羅することができなかったことが理由であると考えられる．しかし，表情の特徴についてはテストデータであっても良好に再現できており，提案手法によって被験者の特徴に影響されない表情特徴の抽出を実現していることが伺える．

図 5.5: 学習画像における表情入れ替え画像生成結果

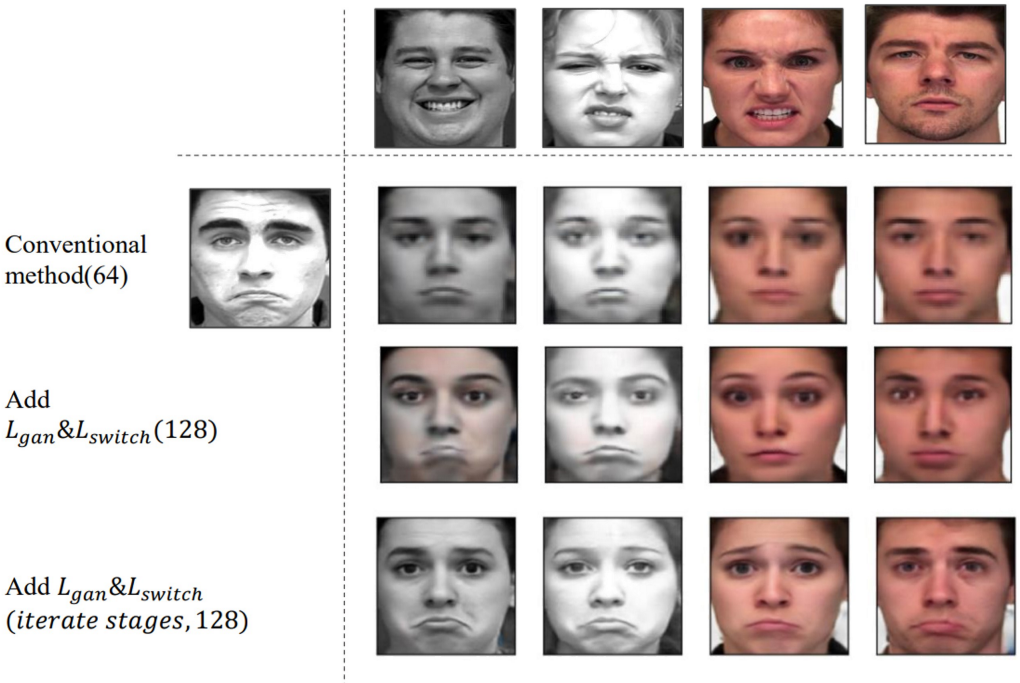

図 5.6: テスト画像における表情入れ替え画像生成結果

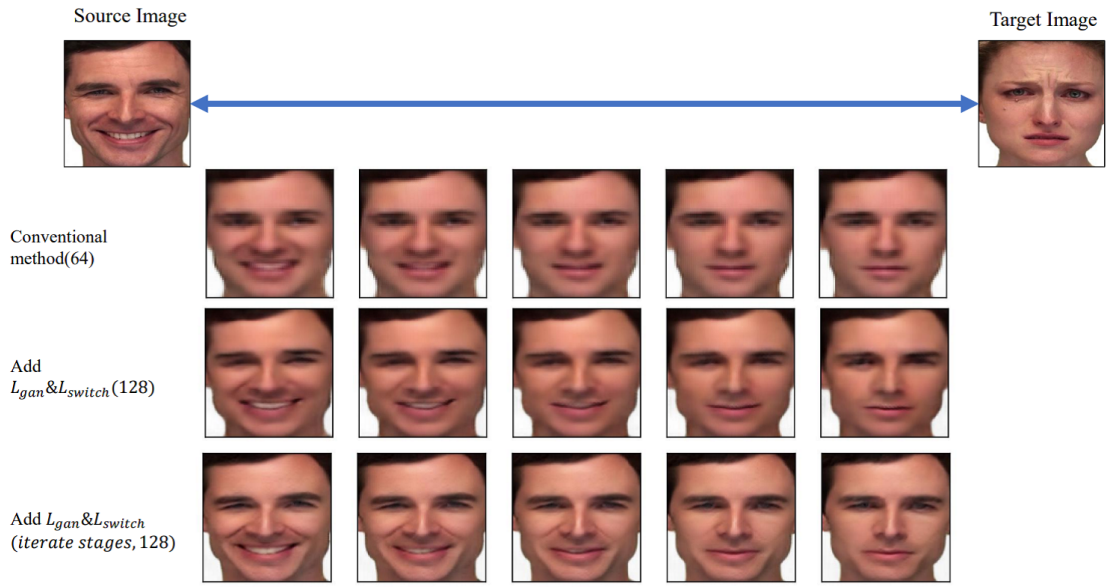

図 5.7: 中間表情生成結果 1

獲得された表情特徴の連続性の評価を目的として，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_{src} と z_{trg} の間を線形補間することで中間表情の画像生成を行う．仮に，モデルの潜在変数として連続的な表情特徴が獲得されていれば，生成される画像の表情は z_{src} の表情から z_{trg} の表情へと，表情や被験者の特徴が崩れることなく，徐々に変化することが期待される．

図 5.7 は被験者特徴をソース画像とした場合における中間表情の生成結果である．最も上段の画像はソース画像とターゲット画像を表しており，2 段目以降は各手法によって生成された中間画像を表している．まず先行手法の結果に注目すると，ある程度は中間表情の生成が行えており，徐々に表情が変化する様子も確認できるが，やはり生成画像がぼやけてしまっており特に中間画像ではしわなどの要素は失われてしまっていることが確認できる．提案手法では，中間の表情であっても鮮明な画像が生成できており，しわの濃さなども段階的に変化していることが確認できる．また，生成された中間表情は意味的にも中間な表現になっていることが分かる．つまり，図 5.7 では喜びと悲しみの中間の画像は悲しそうな笑顔となっている．このように，補間によって生成された画像が鮮明さを保ったまま連続的に表情の変化している点から，ExpressionEncoder の潜在変数として獲得された特徴は，表情の連続性を維持した表情特徴であることが確認できた．

図 5.8 は表情特徴の target と src，被験者特徴の identity をそれぞれ異なる被験者から抽出し，それぞれの手法によって中間画像の生成を行った結果を表している． 図 5.7 で結果を示した実験と大きく異なる点は，被験者特徴を表情特徴を抽出した被験者と異なる被験者から抽出しているため，ここで生成される画像には再構築は含まれず，すべての表情が被験者に対して存在しない表情である点である． 中間画像の生成結果に注目すると，図の結果と同様に従来手法ではぼやけてしまっており、詳細な表情の特徴は確認できない．また，繰り返し学習を行っていない提案手法についても，生成画像が不鮮明であったり，画像が崩れていしまっていることが確認できる．一方提案手法では，表情の埋め込みおよび中間表情の生成を鮮明に，かつ被験者特徴と表情特徴を崩すことなく行えている． この結果から，提案手法を用いることで被験者に対して存在しない中間表情の生成が可能であることが示されるとともに，表情特徴を被験者特徴と分離した状態かつ，表情の連続性を保った特徴として獲得できていることが確認された．

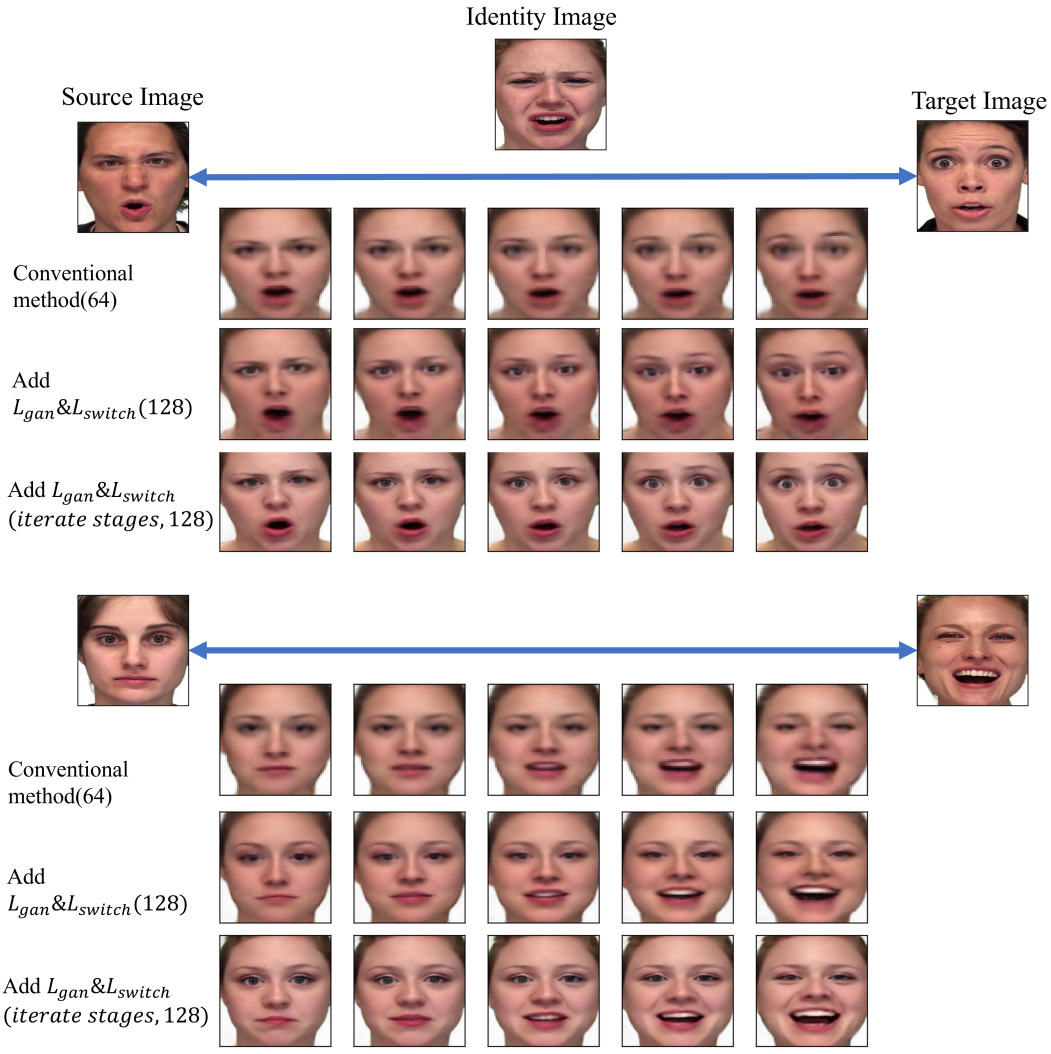

図 5.8: 中間表情生成結果 2

5.4 まとめ

本章では，前章で提案した手法に対して改良を行い，顔のしわなど表情を表現する上で重要な細かな要素を捉えた，より効果的な表情特徴の獲得を行った．

前章で提案した手法では，VAE の特性から生成画像がぼやけてしまっており，表情の細かな要素を表情特徴賭して獲得できていないことが示唆されていた．そこで，本章で提案した手法では生成画像の鮮明化を目的とした損失関数を学習に導入するとともに，表情特徴が被験者間で共通のモノ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で，被験者特徴と表情特徴を分離しつつ，より効果的な表情特徴の獲得を実現した．さらに，学習ステップを先行手法の 1 巡のみ行う方式から繰り返し行う方式に変更することで，異なる目的を持つ 2 つの Encoder が互いに影響を与え合いながら学習することを可能にし，被験者特徴と表情特徴の分離度を高めることを実現した．

実験では，潜在空間利用した表情認識によって表情特徴が被験者に依存しない特徴として獲得されていることを示し，画像生成では被験者特徴と表情特徴の分離度，および表情特徴の連続性を評価した．また，それにより提案手法によって獲得された特徴の有効性を示した．

第6章　結論

6.1　本論文で得られた成果および課題

本論文では，機械学習において表情を扱う際に生じる「表情の連続性の欠落」に対して問題を提起し，表情の連続性を機械学習で取り扱うための手法の提案を行った．各章で得られた成果および課題は以下の通りである．

静止画像を対象とした表情認識モデルの動画像への効果的な適用手法

静止画像を用いて学習を行ったモデルを効果的に動画像へ適用することを目的として，Attention機構を導入し時間方向の表情の連続性を考慮する手法の提案を行った．静止画像を対象としたモデルをそのまま動画像に適用すると，隣接したフレーム間の微妙な表情変化によって認識結果が大きく変動してしまう問題があった．そこで提案手法では，静止画像を対象とした学習済みモデルに対して，時間方向の特徴量に対して重みづけを行うネットワークを追加で導入し，さらに動画像を利用した損失関数によって追加ネットワークを学習することで，表情の連続性をモデルが考慮しそのような隣接フレーム間の認識結果のブレを抑制する．実験では，動画像に対する提案手法の有効性を確認し，静止画像を対象としたモデルの認識対象を動画像に拡張することを実現した．

今後の課題は，照明変化や顔の向きの変化などにより，静止画像による認識結果がより大きく変動する動画像データに対する提案手法の有効性を検証する必要がある．また，モデル構造の検討や，提案手法を既存の学習済みモデルに導入した結果についても確認を行い，提案手法の精度向上や適用可能な対象についても検証を行っていく必要がある．

表情ラベルを利用しない連続的な表情特徴の獲得手法

アノテーションによる表情の離散化を解消することを目的として，感情ラベルなどの表情に関連する情報を学習に利用することなく連続的な表情特徴をモデルの潜在変数として獲得する手法の提案を行い，その有効性を検証した．提案手法はVAEの枠組みを拡張し，被験者IDを利用した2段階の学習により，被験者特徴と表情特徴をそれぞれ異なるEncoderの潜在変数として獲得する．従来の機械学習で表情を扱う手法では，表情に対してアノテーションを行うため，ラベルが付与されることによる表情の離散化や人の主観性が介入することによる教師信号の曖昧性などが課題であった．提案手法では，そのような表情に対するアノテーションを行わずに，顔画像から被験者特徴と表情特徴を分離した状態で抽出することで，表情本来の連続性を維持した特徴の獲得を可能にする．提案手法によって獲得された特徴を潜在空間の可視化や画像生成，ユークリッド距離を用いた表情認識によって評価した結果，連続的な表情特徴がモデルの潜在変数として獲得されていることが確認された．

課題として，提案手法によって生成される画像がぼやけており不鮮明であることから，詳細な表情の特徴が獲得できていないことが挙げられ，次章においてこの課題に対して手法の提案を行っている．

詳細な要素を捉えた連続的な表情特徴の獲得手法

しわなどより詳細な表情の要素を捉えた連続的な表情特徴の獲得を目的として，前章で提案した手法に対して改良を行い，その有効性を検証した．提案手法で行った改良は，2 種類の損失関数の導入と学習ステップの変更である．まず，損失関数の導入では，画像の鮮明化を目的とした損失関数と鮮明化による過学習を抑制する損失関数を導入し，画像の鮮明化を行いながら表情特徴が被験者特徴と絡み合った状態で獲得されることを抑制した．また，学習ステップについては，先行手法の 1 巡のみ行う方式から繰り返し行う方式に変更することで，異なる目的を持つ 2 つの Encoder が互いに影響を与え合いながら学習することを可能にし，被験者特徴と表情特徴の分離度を高めることを実現した．これらの改良により，提案手法は画像生成・表情認識のタスクにおいて先行手法や従来手法よりも良好な結果を示し，しわなどより詳細な表情の要素を捉えた表情特徴をモデルの潜在変数として獲得していることが確認された．

今後の課題として，獲得された特徴を利用して，ファジーな表情認識を可能にする表情認識手法の検討や，提案手法で生成された画像をデータ拡張として利用した場合の表情認識精度の変化などの検証が必要である．また，データセットを拡充し，学習に利用できる被験者数や表情のバリエーションが増えた場合に，提案手法によってより表現力の高い表情特徴の獲得が行えるかについても検証を行っていく必要がある．

以上のように，本論文では機械学習を利用して表情を扱う際に失われることの多い表情の連続性に対して，それらを考慮する手法の提案を行った．内容としては第 3 章の認識に対する連続性から，第 4 章，第 5 章の獲得される表情特徴の連続性に対して手法の提案・考察を行い，その有効性の検証を行った．

謝辞

本研究を進めるにあたり終始多大なるご指導とご助言，素晴らしい研究環境を賜りました長尾智晴先生に深く感謝致します．また，本論文をまとめるにあたり，貴重なご指導ご助言をいただきました森辰則先生，四方順司先生，富井尚志先生，白川真一先生に感謝申し上げます．

本研究をの遂行にあたりご支援いただきました長尾研究室の皆様，特に多岐にわたりご助言・ご意見頂きました小林雅幸様に感謝申し上げます．

最後に，社会人でありながら研究の道に進むことを後押ししてくれた両親と友人に心からの感謝を伝え，謝辞および本論文の締めとさせていただきます．

参考文献

[1] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7132–7141, 2018.

[2] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. Residual attention network for image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3156–3164, 2017.

[3] Paul Ekman and Wallace V Friesen. Constants across cultures in the face and emotion. *Journal of personality and social psychology*, Vol. 17, No. 2, p. 124, 1971.

[4] Hong-Wei Ng, Viet Dung Nguyen, Vassilios Vonikakis, and Stefan Winkler. Deep learning for emotion recognition on small datasets using transfer learning. In *Proceedings of the 2015 ACM on international conference on multimodal interaction*, pp. 443–449. ACM, 2015.

[5] Samira Ebrahimi Kahou, Xavier Bouthillier, Pascal Lamblin, Caglar Gulcehre, Vincent Michalski, Kishore Konda, Sébastien Jean, Pierre Froumenty, Yann Dauphin, Nicolas Boulanger-Lewandowski, et al. Emonets: Multimodal deep learning approaches for emotion recognition in video. *Journal on Multimodal User Interfaces*, Vol. 10, No. 2, pp. 99–111, 2016.

[6] Shichuan Du, Yong Tao, and Aleix M Martinez. Compound facial expressions of emotion. *Proceedings of the National Academy of Sciences*, p. 201322355, 2014.

[7] Ronak Kosti, Jose M Alvarez, Adria Recasens, and Agata Lapedriza. Emotion recognition in context. In *The IEEE Conference on Computer Vision and Pattern Recognition*, Vol. 1, 2017.

[8] Albert Mehrabian. Framework for a comprehensive description and measurement of emotional states. *Genetic, social, and general psychology monographs*, 1995.

[9] Paul Ekman and Wallace V Friesen. *Facial action coding system: A technique for the measurement of facial movement*. Consulting Psychologists Press, 1978.

[10] Timur Almaev, Brais Martinez, and Michel Valstar. Learning to transfer: transferring latent task structures and its application to person-specific facial action unit detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3774–3782, 2015.

[11] Kaili Zhao, Wen-Sheng Chu, and Honggang Zhang. Deep region and multi-label learning for facial action unit detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[12] Andrés Romero, Juan León, and Pablo Arbeláez. Multi-view dynamic facial action unit detection. *Image and Vision Computing*, p. 103723, 2018.

[13] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pp. 3104–3112, 2014.

[14] Thang Luong, Hieu Pham, and Christopher D. Manning. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 1412–1421, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[16] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. Vol. abs/1312.6114, , 2013.

[17] Diederik P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. Semi-supervised learning with deep generative models. In *Advances in neural information processing systems*, pp. 3581–3589, 2014.

[18] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in neural information processing systems*, pp. 2172–2180, 2016.

[19] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. 2017.

[20] Yann LeCun and Corinna Cortes. MNIST handwritten digit database. 2010.

[21] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[22] Luan Tran, Xi Yin, and Xiaoming Liu. Disentangled representation learning gan for pose-invariant face recognition. In *CVPR*, Vol. 3, p. 7, 2017.

[23] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, Vol. 27. Curran Associates, Inc., 2014.

[24] Yu Liu, Fangyin Wei, Jing Shao, Lu Sheng, Junjie Yan, and Xiaogang Wang. Exploring disentangled feature representation beyond face identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2080–2089, 2018.

[25] Diane Bouchacourt, Ryota Tomioka, and Sebastian Nowozin. Multi-level variational autoencoder: Learning disentangled representations from grouped observations. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[26] Ian J Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, et al. Challenges in representation learning: A report on three machine learning contests. In *International conference on neural information processing*, pp. 117–124. Springer, 2013.

[27] Niki Aifanti, Christos Papachristou, and Anastasios Delopoulos. The mug facial expression database. In *Image analysis for multimedia interactive services, 2010 11th international workshop on*, pp. 1–4. IEEE, 2010.

[28] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[29] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pp. 1026–1034, 2015.

[30] Patrick Lucey, Jeffrey F Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *Computer Vision and Pattern Recognition Workshops, 2010 IEEE Computer Society Conference on*, pp. 94–101. IEEE, 2010.

[31] Steven R Livingstone, Katlyn Peck, and Frank A Russo. Ravdess: The ryerson audio-visual database of emotional speech and song. In *Annual meeting of the canadian society for brain, behaviour and cognitive science*, pp. 205–211, 2012.

[32] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pp. 91–99, 2015.

[33] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[34] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, Vol. 9, pp. 2579–2605, 2008.

[35] David Arthur and Sergei Vassilvitskii. K-means++: The advantages of careful seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 1027–1035, 2007.

[36] L. McInnes, J. Healy, and J. Melville. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *ArXiv e-prints*, February 2018.

[37] Durk P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. Semi-supervised learning with deep generative models. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, Vol. 27. Curran Associates, Inc., 2014.

［38］ Yoshihisa Kanou and Tomoharu Nagao. Separation of the latent representations into" identity" and" expression" without emotional labels. In *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pp. 1638–1644. IEEE, 2020.

研究業績

論文誌

狩野悌久，長尾智晴：弱教師あり学習による連続的な表情特徴の獲得，情報処理学会論文誌：数理モデル化と応用 (TOM), vol.15, No.2, pp.11-20, 2022.

国際会議発表

Yoshihisa Kanou, and Tomoharu Nagao: Separation of the Latent Representations into ”Identity” and ”Expression” without Emotional Labels, Proceedings of 2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC2020), pp.1638-1644, 2020.

国内学会発表

狩野悌久，長尾智晴 : Attention 機構を用いた深層学習による表情認識，情報処理学会 第 81 回全国大会, 2019.

造酒裕貴，狩野悌久，長尾 智晴：深層強化学習を利用した株式売買戦略の構築，第 28 回インテリジェント・システム・シンポジウム (FAN2018), 2018.