あなた専用のお手本映像で上達支援

近年の深層学習の発展は著しく，少し前までは想像もつかなかったことが実現されつつあります．メディア生成はその代表例であり，世界中で驚くような研究成果が報告され始めています．本稿では，深層学習によるメディア生成の研究動向，およびメディア生成のアプリケーションの中で着目している「上達支援のためのメディア生成」を題材に，従来技術の課題とNTTコミュニケーション科学基礎研究所の取り組みについて紹介します．

金子 卓弘（かねこ たくひろ）／平松 薫（ひらまつ かおる）
柏野 邦夫（かしの くにお）

NTTコミュニケーション科学基礎研究所

深層学習の広がり

近年，深層学習と呼ばれる技術の著しい発展により，少し前までは想像もつかなかったことが実現されつつあります．メディア生成はその代表例であり，世界中で驚くような研究成果が報告され始めています．例えば，手に入れたい画像があったときに，テキストを入力すれば自動的に写真のような画像を新規に創出・生成してくれる技術[(1)]，自分の持っている画像を有名画家が描いたような特徴的な絵画に自動的に変換してくれる技術[(2)]などがあります．

数年前までは，画像識別や音声認識など比較的正解が明確な問題に対して精度を向上させる研究が主でしたが，近年は，深層学習ならではの表現能力の高さを活かし，新たな概念・モデル構造を導入することで，より複雑な問題も解決できるようになってきています．特に，メディア生成の場合は，先の例のように，あるものをつくりたいときの創作補助やあるものに変換したいときの加工補助など，人々のさまざまな願望を具現化しサポートする技術として期待が高まっています．

上達支援のためのメディア生成

「ボールを速く投げたい」「英語をきれいに発音したい」「印象の良い表情にしたい」など「何かをしたい」というときに，「どうすれば良いのかやり方が分からない…」といった経験は誰しもがしたことがあるのではないでしょうか．このようなときに有効な方法としては，詳しい人に教えを乞うたり，自身で書物やインターネットを頼りに情報を探したりといった方法が考えられます．前者は個人に合った具体的な指示を得られる点，後者は自己解決できる点で有用ですが，一方で，前者は適切な人を探し出すことが難しく，後者は特定の個人を対象にした情報ではないため自身のケースにうまく当てはまるものを見つけ出すことが難しいという問題があります．このような問題に対し，私たちは前者の利点である「個人性」「具体性」と後者の利点である「自己解決性」の両者を満たすものとして「あなた専用のお手本映像の生成」システムの実現をめざしています．

あなた専用のお手本映像の生成

提案する「あなた専用のお手本映像の生成」システムのコンセプトを**図1**

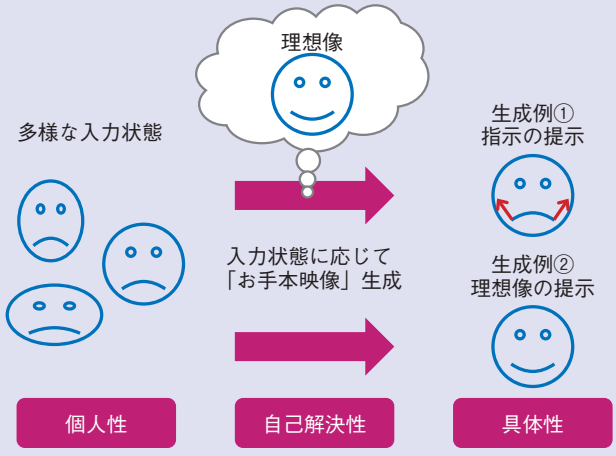


図1　「あなた専用のお手本映像の生成」システム（例：顔の表情改善の場合）

に示します．本システムでは，ユーザが与えたメディア情報（画像や音声など）を基に解析を行うことで「個人性」を実現します．次に，出力として具体的な指示や理想像を提示することで「具体性」を実現します．さらに，これらの処理を一貫して自動的に行うことで「自己解決性」を実現します．

NTTコミュニケーション科学基礎研究所が考案した表情改善のためのフィードバックシステム[(3)]の動作例を**図2**に示します．本システムでは，前準備として，事前に収集した学習用データを基にフィードバックのルールを学習します．この学習用データには不特定多数の人物の顔画像があれば良く，学習段階でユーザ本人の顔画像は不要であり，さらに，学習データ中の特定の人物に対して複数の表情の顔画像を収集する必要もありません．このため，ユーザに対してデータ収集の負荷をかけることなく，さらに，高いデータ収集コストを要することなく，「自己解決性」を実現することができます．フィードバックを生成する際には，ユーザの顔表情をカメラで撮影し現状把握を行います．この情報を基に解析を行うことで各ユーザの現状に最適な理想状態を「個人性」を考慮して求めます．そして，この解析結果を基にどの顔パーツをどのぐらい修正すれば良いかを矢印で提示することで「具体性」のある情報提示を行います．具体的な指示が与えられているので，ユーザはこの指示に従って顔パーツを動かすことで理想状態に近づくことができます．ここでは，表情改善を例に説明しましたが，同様の方式は，上手な話し方の練習など，音声に対しても適用することができます．

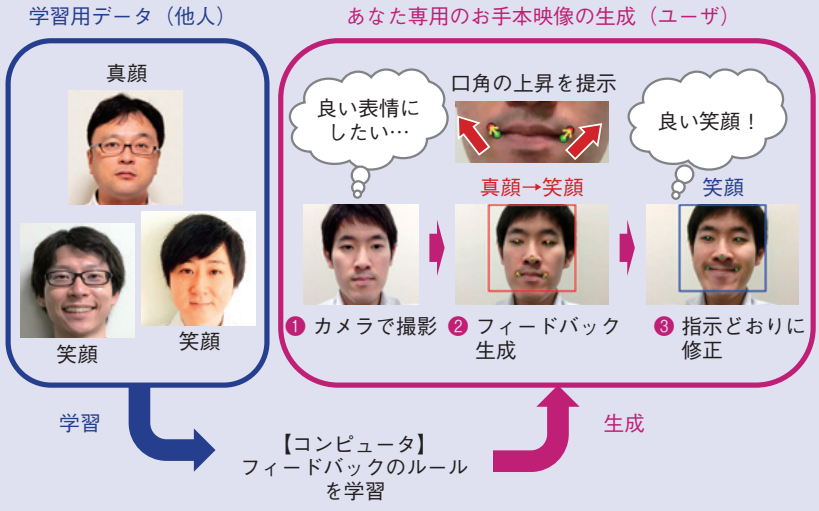


図2　表情改善のためのフィードバックシステムの動作例

多様な入力状態への対応と具体的な情報提示

前述のようなシステムを実現するためには，2つの技術的な課題がありました．1番目の課題は多様な入力状態への対応です．本システムでは人を対象にしていますが，人は老若男女千差万別であり，最適なフィードバックも人ごとに異なります．従来のフィードバックシステムはルールベースで入出力関係を定めるものが主でしたが，この方法ですと多様な入力状態に対応するためには数多くのルールを人手で作成する必要がありました．本研究では，この課題を解決するために学習に基づくアプローチを考案しています．

2番目の課題は入力に対して単に現状把握するだけではなく，それを基に具体的な情報提示を行うことです．従来手法でも多様な入力状態に対応するために学習ベースのアプローチを用いたものがありましたが，これらは現状の判定（笑顔度の算出や表情クラスの識別など）や特徴点の検出（顔パーツの検知など）にとどまっており，具体的なフィードバックを提示するには至りませんでした．この限界の要因としては，提示する情報を具体的に学習するためにはフィードバックの正解

データを大量に集める必要がある点が挙げられます．一般に，フィードバックの正解の付与は専門的な知識を要し，簡単なタスクではありません．このような理由により従来技術では本システムのような学習ベースのフィードバック生成システムを実現することは困難でした．

汎用の学習データからフィードバックを学習・生成：Deep Feedback法

2つの課題を解決するために，私たちはフィードバックの正解データを陽に用いることなく，フィードバックのルールを学習・生成することができる新たな深層ニューラルネットワーク（DNN: Deep Neural Network）の仕組みを考案しました．本手法の処理フローを**図3**に示します．まず，モデルの学習時には，顔画像に対して顔パーツ位置と表情クラスが付与されたデータを学習データとして用います．このデータを得るためにも，顔パーツ位置と表情クラスという2つのタスクに対して正解データを用意する必要がありますが，これらのタスクはフィードバックの正解を付与することと比較すると簡易であり，クラウドソーシングなどを用いて比較的容易に大量のデータを収集することができます．この学習データを用いて，2つのタスクに対して共通する特徴量空間を持つDNN（マルチタスクDNN）を学習します．この学習により，マルチタスクDNNは共有する特徴量空間内で入力画像，顔パーツ位置，表情クラスの3つの関係を獲得することができます．

フィードバックを生成する際は，この性質を活用して，新たに開発したマルチタスクDNNからフィードバックを引き出す方法（Deep Feedback法）を用います．

① 入力画像を用いてDNNの各層の特徴量の初期値を求め，現状の顔パーツ位置・表情クラスの推定を行います．

② 表情クラスの出力部分に対して，目標状態に到達するまでの不足量を算出します．

③ ②の不足量を低減するために，DNNの中の所定の層の特徴量を誤差逆伝播法で更新を行います．一般的なDNNの学習では特徴量の値は固定してモデルパラメータを更新しますが，このDeep Feedback法ではモデルパラメータは固定して特徴量の値を更新する点がポイントです．

④ 更新した特徴量を基に順伝播法を用いることによって顔パーツの位置を更新します．

最後に，更新前と更新後の顔パーツ位置の差分を矢印として提示することでフィードバックを与えることができます．

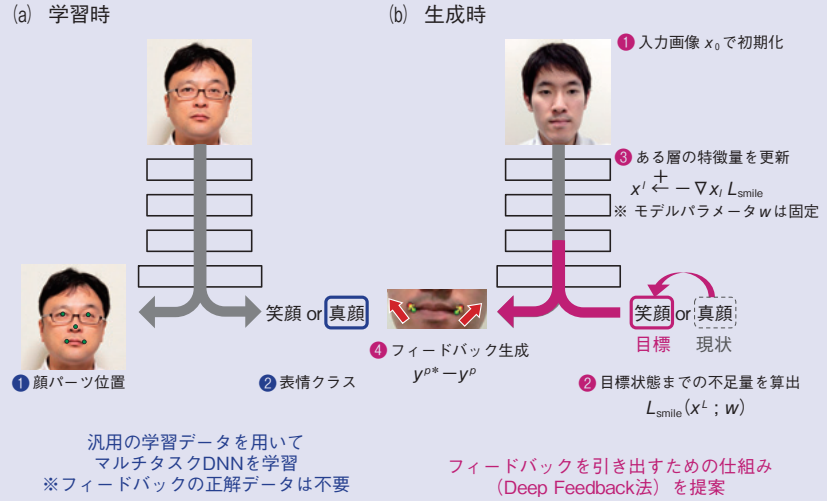


図3　汎用の学習データからフィードバックを学習・生成：Deep Feedback法

インタラクティブな操作による理想像の生成・探索

前述の枠組みでは，ユーザにとって理想となる笑顔をコンピュータが1

つ見つけ出し提示しますが，笑顔と一口に言っても「クスっと笑い」「大笑い」「にやり笑い」などさまざまな笑い方があります．このように正解が複数想定される場合，コンピュータがユーザの意図を一度に完全に把握して最適な理想像を提示することは簡単ではありません．そこで，私たちはユーザがインタラクティブに操作しながら，本当に求めている理想像を探索することができるシステム[4]を考案しました．本システムの概念図を**図4**に示します．本システムでは，ユーザはスライドバーやラジオボタンなど画像編集でよく使われるGUI（Graphical User Interface）を操作しながら，多様な理想像を実際に生成しながら探索することができ，これによってユーザは手軽でかつ直感的に理想像を見つけ出すことができます．

簡易な学習データから笑顔の表現空間を学習：Deep Attribute Controller法

本システムを実現するためには，さまざまな笑顔に対して十分に表現力があり，かつ操作性の高い笑顔の表現空間を獲得することが必要になります．このためのもっとも愚直な方法は，人手で笑顔を分類・整理して空間を構成することですが，笑顔にもさまざまなものが存在し，明確に定義できないものも含まれるため簡単ではありません．そこで，本研究では，笑顔か笑顔でないかという簡易なラベルのついた学習データだけから，表現力と操作性を兼ね備えた笑顔の表現空間を自動的に獲得する方法（Deep Attribute Controller法）を考案しました．具体的には，近年提案されたDNNベースの確率的生成モデルである敵対的生成ネットワーク（GAN：Generative Adversarial Networks）[5]を拡張したモデル（CFGAN: Conditional Filtered GAN）を提案しています．CFGANのモデル構造を**図5**に示します．GANは，乱数を入力として識別器を騙せるくらい真の画像とそっくりな画像を生成する生成器と，真の画像と生成画像を見分ける識別器の2つのネットワークから構成されています．この生成器と識別器がMin-Maxという敵対する条件下で最適化，つまり，生成器はなるべく識別器を「騙せる」ように最適化し，識別器はなるべく生成器に「騙されない」ように最適化することで，真のデータ分布に近いデータを生成できる生成器を得ることを可能にします．提案したCFGANでは，この生成器の入力部分に簡易なラベルの値に応じて潜在データのオンオフを制御するような仕組みを導入し，これによって，高い表現能力と高い操作性を兼ね備えた笑顔の表現空間の学習を可能にしました．本モデルで鍵となるのは，簡易なラベルだけからそのラベルに関して高い表現能力と操作性を持った表現空間を自動的に学習できるようにしたことです．このため，笑顔だけにとどまらず，年齢や髪型などさまざまな属性への応用が可能であり，さらには，画像だけにとどまらず音声などさまざまなメディア情報への応用も期待されます．そのための必須技術として，私たちは肉声と

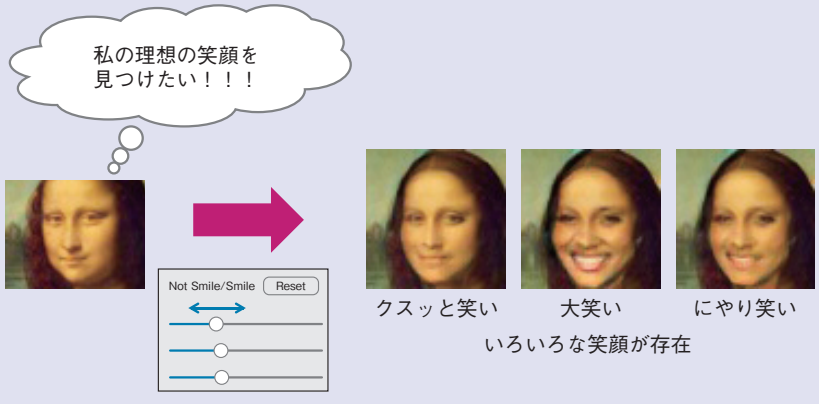


図4　インタラクティブな操作による理想像の生成・探索

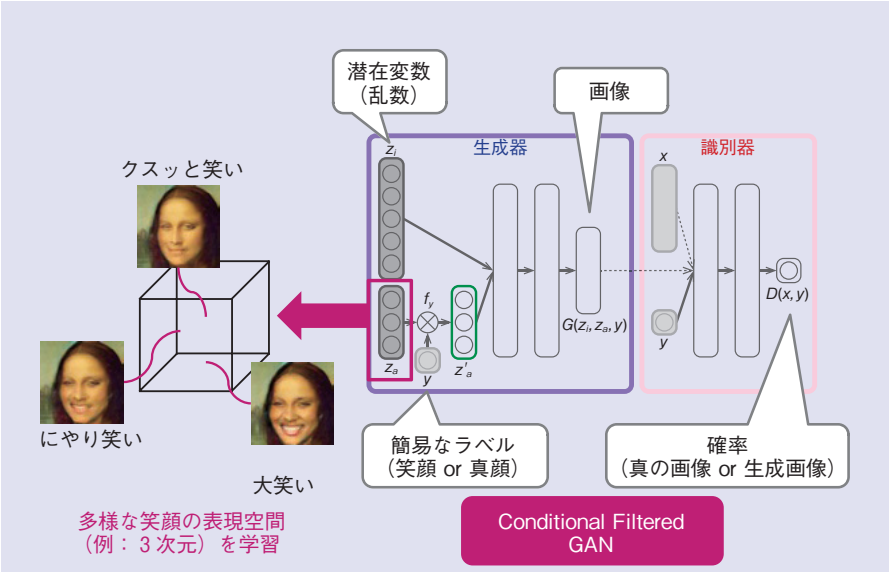


図5　簡易な学習データから笑顔の表現空間を学習：Deep Attribute Controller法

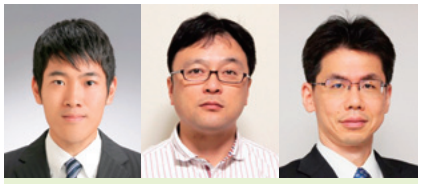
区別のつかない音声の合成[6], [7]や変換[8]のための研究も進めています．

今後の展開

前述のアプローチの鍵は，システムがメディア生成を通していかにユーザに寄り添えるか，ということです．このメディア生成の技術は，上達支援だけにとどまらずユーザが思い描くイメージを具体的なメディア（画像，音声など）として具現化するためになくてはならない技術です．私たちは，どのようなものが生成できたら嬉しいかという想像力と，深層学習を用いたメディア生成に関する知見と経験とを積み重ねながら，将来的には，ユーザのあらゆる願望にこたえられる，極めて高品質なメディア情報を生成できる技術の確立をめざしています．

■参考文献

(1) H. Zhang, T. Xu, H. Li, S. Zhang, X. Huang, X. Wang, and D. Metaxas：“StackGAN: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks,” in arXiv preprint arXiv: 1612.03242, 2016.
(2) L. A. Gatys, A. S. Ecker, and M. Bethge：“Image Style Transfer Using Convolutional Neural Networks,” Proc of CVPR 2016, Los Angels, U.S.A., June 2016.
(3) T. Kaneko, K. Hiramatsu, and K. Kashino：“Adaptive Visual Feedback Generation for Facial Expression Improvement with Multi-task Deep Neural Networks,” Proc. of ACMMM 2016, Amsterdam, Netherland, Oct. 2016.
(4) T. Kaneko, K. Hiramatsu, and K. Kashino：“Generative Attribute Controller with Conditional Filtered Generative Adversarial Networks,” Proc. of CVPR 2017, Honolulu, Hawaii, July 2017.
(5) I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio：“Generative Adversarial Nets,” Proc. of NIPS 2014, Montreal, Canada, Dec. 2014.
(6) T. Kaneko, H. Kameoka, N. Hojo, Y. Ijima, K. Hiramatsu, and K. Kashino：“Generative adversarial network-based postfilter for statistical parametric speech synthesis,” Proc. of ICASSP 2017, New Orleans, U.S.A., March 2017.
(7) T. Kaneko, S. Takaki, H. Kameoka, and J. Yamagishi：“Generative adversarial network-based postfilter for STFT spectrograms,” Proc. of Interspeech 2017, Stockholm, Sweden, August 2017.
(8) T. Kaneko, H. Kameoka, Hiramatsu, and K. Kashino：“Sequence-to-sequence voice conversion with similarity metric learning using generative adversarial networks,” Proc. of Interspeech 2017, Stockholm, Sweden, August 2017.



（左から）金子　卓弘／平松　薫／
柏野　邦夫

深層学習の著しい発展により，少し前までは想像もつかなかったことが実現されつつあります．メディア生成はその代表例であり，私たちは，ユーザのあらゆる願望にこたえられる，極めて高品質なメディア情報の生成技術の確立をめざしています．

◆問い合わせ先
NTTコミュニケーション科学基礎研究所
メディア情報研究部
TEL　046-240-3708
FAX　046-240-4708
E-mail　kaneko.takuhiro@lab.ntt.co.jp