あなた専用のお手本映像で上達支援

深層学習によるメディア生成の可能性

Generative personal assistance with audio and visual examples

Deep learning opens the way to innovative media generation

メディア情報研究部

金子　卓弘 Takuhiro Kaneko

プ ロ フ ィ ー ル

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 研究員。2012年東京大学工学部卒業。2014年同大学大学院情報理工学系研究科修士課程修了、2017年博士課程入学。2014年日本電信電話株式会社入社。コンピュータビジョン、画像処理、音声信号処理、機械学習の研究に従事。2012年日本機械学会畠山賞受賞。同年ICPR2012 Best Student Paper Award受賞。電子情報通信学会会員。

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、例えば、任意の乱数から本物と見分けのつかないような画像を生成する技術など、世界中で驚くような研究成果が報告され始めています。同じ深層学習でも、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。

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」という時に、「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。このような時には、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よくとられます。前者は個人に合った具体的な指示が得られる点、後者は自己解決できる点で有用ですが、一方で、前者は詳しい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ものではないため自身のケースにうまく当てはまる情報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」システムの実現を目指しています。

あなた専用のお手本映像の生成

図1に提案システムのコンセプトを示します。このシステムでは、ユーザが与えたメディア情報（画像や音声）を元に解析を行うことで「個人性」を実現し、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、さらにこれらの処理を自動的に行うことで「自己解決性」を実現します。図2に、実際に私たちのグループが考案した表情改善のためのフィードバックシステム[1]の動作例を示します。本システムでは、まず、ユーザの顔表情をカメラで撮影して現状把握を行います。次に、この情報を元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求め、どの顔パーツをどのぐらい修正すればよいかということを矢印で提示します。具体的な情報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。同じような方式は、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。

多様な入力状態への対応と具体的な情報提示

上記のようなシステムを実現するためには、二つの点で技術的な難しさがありました。一つ目は多様な入力状態への対応です。このシステムが対象としている人は千差万別であり、最適なフィードバックも個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、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。二つ目の難しさは、入力に対して単に認識するだけではなく、それを元に具体的な情報提示を行う点です。従来の画像認識技術では、クラス（例えば表情）の識別や特徴的な点（例えば顔パーツ）の検出などを行うことはできましたが、提示すべき情報の生成は困難でした。私たちは、この問題を解決するために、深層学習における新たな情報伝播方法を考案し、多様な入力状態の理解と具体的な情報提示とを一体的なモデルで実現することを可能にしました。

深層学習によるメディア生成の今後の展望

上記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。そのための必須技術として、私たちは、インタラクティブに操作しながら理想像を探索・提示する方法[2]、肉声と見分けのつかない音声の合成方法[3]なども考案してきています。これらは、上達支援だけにとどまらずユーザが思い描くようなメディアを生成するために無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、ユーザのあらゆる要望に応えられる、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目指しています。

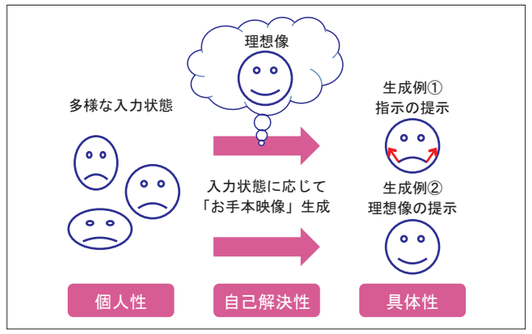

図1:「あなた専用のお手本映像の生成」システム
（例:顔の表情改善の場合）

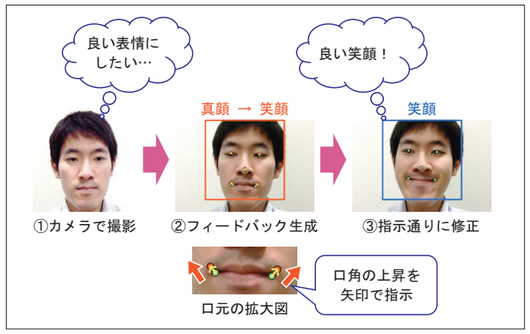

図2:表情改善のためのフィードバックシステムの動作例

関 連 文 献

[1] T. Kaneko, K. Hiramatsu, K. Kashino, “Adaptive visual feedback generation for facial expression improvement with multi-task deep neural networks,” in *Proc. 24th ACM International Conference on Multimedia (ACMMM)*, pp. 327-331, 2016.
[2] T. Kaneko, K. Hiramatsu, K. Kashino, “Generative attribute controller with conditional filtered generative adversarial networks,” in *Proc. 30th IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017 (to appear).
[3] T. Kaneko, H. Kameoka, N. Hojo, Y. Ijima, K. Hiramatsu, K. Kashino, “Generative adversarial network-based postfilter for statistical parametric speech synthesis,” in *Proc. 42nd IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4910-4914, 2017.