2R-07

音響情報を用いた一枚画像からの動画生成

土屋 志高† 板摺 貴大† 夏目 亮太† 加藤 卓哉† 山本 晋太郎† 森島 繁生‡
早稲田大学† 早稲田大学理工学術院総合研究所‡

1. はじめに

人間は，人の声や海のさざ波のような聴覚情報から，人が話す様子や波が動く様子のような視覚情報を想像することが可能である．このような能力をコンピュータで実現する研究が近年提案されている[1][2]．Suwajanakornら[1]は，音声から RNN により口形を予測し，テクスチャを合成することで顔の動画を生成した．Shlizermanら[2]は，ピアノやヴァイオリンなどの楽器演奏からLSTMにより人間の腕や指などのボーンを予測し，アバターの動きを動画で生成した．これらの手法では口の特徴点や人間のボーンの情報が必要となるため，特徴点が存在しない音と動きが連動した現象に対しては適用できないという問題点がある．

本稿では，動かしたい対象の画像と音から音と動きのタイミングが合う動画の生成を行う．その際に，特定の対象に特化した特徴量を用いず画像と音を入力とすることで，任意の対象に対して学習し動画を出力することが可能となる．

2. 提案手法

本研究では，数秒の音から得られる特徴と一枚の画像から GAN を用いて動画を生成する．提案手法におけるネットワークを図 1 に示す．Bidirectional LSTM (BLSTM) により音の時間的な変化の特徴を抽出する．得られた特徴量と画像から GAN の Generator (G) を用いて画像を生成する．生成された連続した画像をまとめて GAN の Discriminator (D) により時間的に自然な画像を生成するように学習を行う．

2.1 音の時間的変化の特徴抽出

音のタイミングに合う動画を生成するためには動画の各フレーム画像に対して音の特徴量を対応付けることが必要になる．音が44.1 [kHz]，画像が30 [fps]で構成される動画では$30/44100$秒ごとにフーリエ変換することで，フレーム毎に735次元のスペクトログラムSPを得る．得られたSPから BLSTM により，各フレーム$t(1 \leq t \leq T, T$は動画のフレーム数)に対してk次元の特徴量SB_tを得る．BLSTM は時間的な変化の特徴を得られるため，音が鳴っていない時間がある場合でもそのフレームでの音の特徴量を抽出できる．

2.2 画像と音からの画像生成

画像はフレーム毎に生成する．tフレーム目のSB_tを入力画像Iのサイズ$H \times W$になるようにタイルしたものをSB_t^*とする．IとSB_t^*を結合することで $(k+3)$チャンネル，$H \times W$のテンソルIS_tを得る． IS_tから tフレーム目の画像IG_tを生成する．Iの見た目を保持した画像を生成するためにGとして U-Net[3]を用いた．Tフレームの生成された画像IGを画像一枚ごとではなく，TフレームまとめてDに通すことで画像の連続性を考慮した自然な出力動画を生成するように学習を行う．

3. 実験

3.1 実験設定

音に無関係の動きを排除するために定点カメラでの動画を学習に用いた．学習には縦横64ピクセルの画像を用いた．SB_tの次元数は$k = 32$とした．また，入力音の長さは4秒とし，$T = 120$とした．

3.2 データベース

提案ネットワークの学習時には，入力画像I，入力音S，Tフレームの生成画像IGに対する正解画像が必要になる．データベースの動画をランダムな区間で4秒ごとに区切り，4秒の動画を120枚の画像と音に分離する．この120枚の画像を正解画像とし，120枚からランダムに一枚を選択し，それを入力画像とする．今回は人の口，手，海，花火の動画を用いてそれぞれ別々に学習した．口，手の動画は筆者らが撮影したもの，海，花火は YouTube に公開されている動画を使用した．学習に用いたそれぞれの動画の本数と総秒数を表 1 に示す．

表 1 各対象の動画の本数と総秒数

	口	手	海	花火
動画の本数	6	30	6	4
総秒数[s]	408	944	10560	3904

“Video Synthesis from Sounds and A Single Image”
†Yukitaka TSUCHIYA †Takahiro ITAZURI
†Ryota NATSUME †Takuya KATO
†Shintaro YAMAMOTO ‡Shigeo MORISHIMA
†Waseda University
‡Waseda Research Institute for Science and Engineering

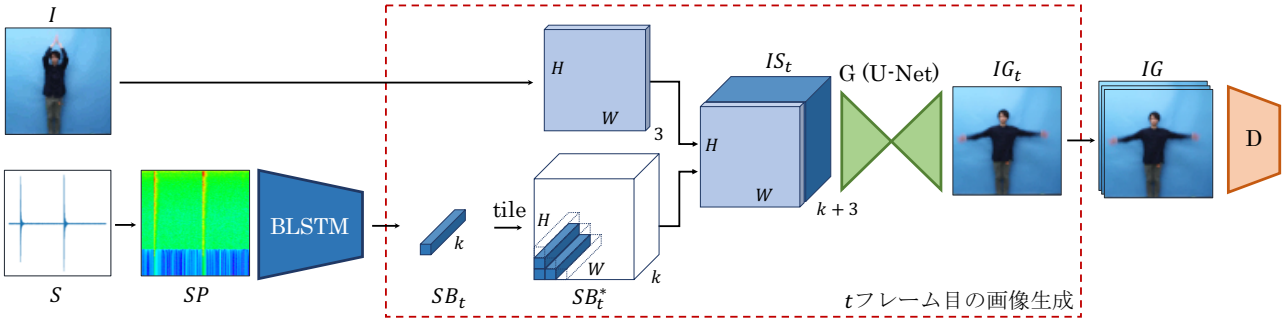

図 1　提案ネットワーク

4. 結果と考察

学習時には含まれない画像と音を入力とした時の人の口，手，海[a]，花火[b]の動画生成の結果を図 2 に示す．入力音を抽出した元の動画の中から無作為に画像一枚を選択し，入力画像とした．

図 2 に示す通り，いずれの入力画像に対してもその外見を保持したまま，音のタイミングに合う動画の生成に成功した．このことから提案ネットワークは特定の対象に限定することなく，任意の対象に対して学習し，動画を生成できることが確認された．

口の 120 フレーム目は，生成画像では口を開けているが正解画像は口を閉じている．これは学習時に声を発した後にも口を数秒間開けているという動画が含まれていることによるものであると考えられる．また，海の 90 フレーム目は生成画像では白波があるが正解画像には白波がない．これは元の動画では画面の外の波の音が入力音に含まれており，その音のタイミングに合うように画像を生成しているため，白波が画像として生成されたと考えられる．

以上のように生成動画と元の動画では動きの変化の仕方は必ずしも同じではないことがわかる． 本研究においては元の動画と生成動画が同じである必要はなく，入力音に対して入力画像の動きの変化が自然であることが重要である．

5. まとめと今後の課題

本研究では，音と動きが連動している対象に対して一枚の画像と数秒の入力音から，画像の見た目を保持した音のタイミングに合う動画を生成する手法を提案した．4 種類の対象に対して実験を行い，入力音のタイミングに合う動画を生成した．本稿では，動画内で音を発する対象が一つに限定されていることや，音に無関係な動きを排除するために定点カメラの動画が必要になる．今後の課題として，一枚の画像の中に音に関係するものが複数存在する場合や，カメラが動く場合の動画での動画生成がある．

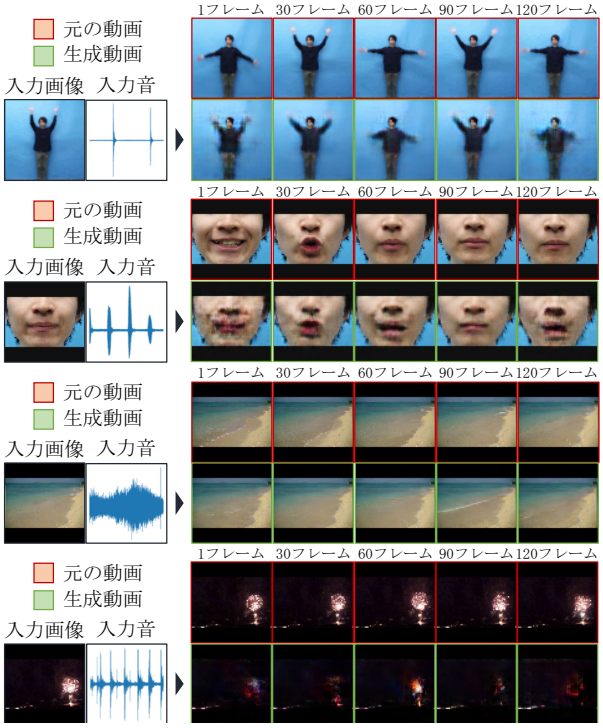

図 2 各対象の元の動画と生成動画の比較

謝辞

本研究の一部は，JST ACCEL (JPMJAC1602) の支援を受けた．

参考文献

[1] S. Suwajanakorn et al. "Synthesizing obama: learning lip sync from audio." SIGGRAPH 2017.
[2] E. Shlizerman et al. "Audio to body dynamics." CVPR 2018.
[3] O. Ronneberger et al. "U-Net: Convolutional Networks for Biomedical Image Segmentation" MICCAI 2015.

[a] https://youtu.be/3OBbYfgelkQ
[b] https://youtu.be/a4i_Pr_fB3I