4X-07

三次元顔特徴点を用いたDNNに基づく 2Dフォトリアリスティック顔動画像生成の検討

佐藤　一樹†　　能勢　隆†　　伊東　燦‡　　伊藤　彰則†
†東北大学大学院工学研究科　　‡東北大学工学部

1　はじめに

近年，より豊かなHuman Computer Interactionの実現のため，リアルな人の姿を持つコンピュータエージェントの実現に向けた人の発話動画像生成に関する研究が数多く行われている．Andersonらは一般的な2次元顔モデルであるActive Appearance Model (AAM)[1]を拡張し，Cluster Adapting Training [2]と隠れマルコフモデルを用いて感情を含む高品質な発話顔動画像の生成手法を提案した[3]．しかしこの手法ではAAMを構築する際に一部手作業で特徴点を指定する必要があるためデータ取得のコストが高いといった問題点がある．Wangらは2Dの顔画像から3Dモデルを生成し，サンプルベースの画像生成と組み合わせる手法を提案した．この手法では再現の難しい口内などをサンプルベースで描画することで自然性の高い3D発話動画を実現しているが，3Dの描画には高いマシンパワーが必要なためモバイル端末などでの使用は難しいと考えられる．

我々はこれまで，Kinect v2を用いることで取得できる顔の各部位の状態を表したAnimation Unit(AU)パラメータを顔画像特徴量とし，隠れマルコフモデル音声合成と同様の枠組みで任意のテキストからAUパラメータを生成し，それをDeep Neural Network(DNN)を用いて画素値系列へと変換することでテキストからの発話顔動画像生成手法を提案してきた[4]．この手法の問題点として，ピクセルベースの生成手法であるために学習データ内に含まれる顔の位置や向きのずれに対応できず，品質の向上に限度があることが挙げられる．そこで我々は新たな手法として，Kinect v2を用いて取得した顔の三次元特徴点を元に顔モデルを構築し，DNNを用いてテキストから顔動画像を生成する手法を提案する．

2　DNNを用いた顔動画像生成

2.1　Kinectで取得した三次元特徴点を用いた顔モデル

本研究では，Kinect v2で取得した顔の三次元特徴点を用いて顔をモデル化する．Kinect v2のHigh definition face tracking API [5]を用いることで，1347点の顔の三次元特徴点を取得することが可能である．取得した三次元特徴点系列のあるフレームに対し，特徴点が構成する各三角形ポリゴンの画素数を決めることでデータ全体における解像度を決定する．ポリゴンの画素数の決定の仕方について記す．n番目のポリゴンの各頂点をそれぞれp_{n1}, p_{n2}, p_{n3}と置くと，このポリゴン内の点pは次の式で表される．

$$p = s(p_{n2} - p_{n1}) + t(p_{n3} - p_{n1}) + p_{n1} \tag{1}$$

ただし$0 \leq s \leq 1,\ 0 \leq t \leq 1, 0 \leq s + t \leq 1$である．$s, t$はテクスチャ座標とみなすことができるため刻み幅$h_n$を導入し$s, t$を$s = lh_n, t = mh_n$と定義し直す．$l, m$は$s, t$の範囲から一意に範囲が決まる．また刻み幅$h_n$によってポリゴンの解像度が決まる．$n$番目のポリゴンの画素数$N_n$と$h_n$の間には式(2)が成り立つから，ポリゴン間の面積比と画素数比が等しいとしてN_nを最小のポリゴンとの比で求めることでh_nを決めることができる．なお本研究では最小のポリゴンの画素数を1とした．

$$N_n = \sum_{k=0}^{\lfloor \frac{1}{h_n} \rfloor} (\lfloor \frac{1}{h_n} \rfloor - k) \tag{2}$$

このようにして決定したあるフレームの各ポリゴンの画素数とその特徴点座標を以後モデルファイルと呼ぶ．モデルファイルに定義された各ポリゴンの画素数を元に，データの各フレームに対してポリゴンのテクスチャを取得するため，全フレームにおいてポリゴンごとの画素数は一定である．テクスチャ座標の各点における画素値は，三次元座標を画像平面に写像して得られた二次元座標に対応する画素値をRGBカラー画像の画素値からバイキュビック補間することで取得する．これを全フレームに対して行うことで，各フレームにおけるポリゴンのテクスチャを得ることができる．本研究では，以上のような三次元特徴点とテクスチャを用いて顔を表現する．

2.2　使用する顔画像特徴量

テクスチャは次元数が大きくそのまま扱うのが難しいため，主成分分析(PCA)を行い次元圧縮したものを特徴量として用いる．次元圧縮する際は，PCAに用いたテクスチャの平均との差分を使用する．またKinectで取得した三次元特徴点は細かく振動しているため，各フレームの三次元特徴点群をモデルファイルの三次元特徴点群へのフィッティングを行うことで振動を除去する．具体的には，発話によってあまり座標が変動しない特徴点に対してフィッティングを行うアフィン変換行列を最小二乗法により求めたあとカルマンフィルタを用いて変換行列の誤差を最小化しアフィン変換を行う．最終的に，得られた三次元特徴点座標とテクスチャのPCA係数をそれぞれの系列の最大値と最小値を用いて0から1の範囲の値をとるように正規化した後一次，二次の動的特徴量をとり，それらをまとめて顔画像特徴量とする．

A study on 2D photo-realistic facial animation generation based on DNN using 3D facial feature points
†Kazuki SATO　†Takashi NOSE　‡Akira ITO　†Akinori ITO
†Graduate School of Engineering, Tohoku University
‡School of Engineering, Tohoku University

表 1: DNN の構造

入力層ユニット数	413	出力層ユニット数	12507
中間層数	3	中間層ユニット数	1024
Optimizer	Adam	活性化関数	tanh
バッチサイズ	100	epoch 数	1000
Dropout 率	0.5		

2.3 DNN の学習及び顔動画像生成の流れ

学習の段階では，収録した音声を用いて生成したコンテキストラベルを入力，顔画像特徴量を出力とし DNN で学習を行う．生成の段階では，任意のテキストから音素継続長を予測したあとそれを元にコンテキストラベルを作成し，DNN に入力することで顔画像特徴量系列を予測する．得られた特徴量系列から顔を復元する．この際生成できるのは顔領域のみになるため，動画として再生すると顔だけが浮かんで喋っているような動画となる．実用を考えるとこれは自然性が低いため，本研究では背景となる動画を用意し，生成された顔動画を背景動画に貼り付けるという手法を用いる．背景動画には学習データと同じように Kinect で収録したものを使用し，収録された三次元特徴点座標を元に，生成された特徴点をアフィン変換，カルマンフィルタを用いてマッピングし，元々の顔の位置に生成した顔を貼り付けることで発話動画を生成する．

3 実験

3.1 実験条件

データセットとして男性話者 1 名が ATR 日本語データベースの文章を 11 文読み上げている様子を Kinect v2 を用いて収録し，学習データとして 10 文，テストデータとして 1 文を用いた．動画のサイズは 400×400 で，音声は 16kHz サンプリングの wav 形式で保存した．PCA には学習データからランダムに 128 フレームを用い，PCA 係数は 128 次元全て使用した．テクスチャは次元数は大きく通常の PCA を用いると計算コストが高いため，実際には Incremental PCA を使用した．コンテキストには，HMM 音声合成の決定木クラスタリングに用いられる先行，当該，後述の音素，アクセント，呼気段落，文長に関する質問への回答を 0 または 1 で表現したベクトル 412 次元と音素内フレーム位置を 0 から 1 の間で表した値の計 413 次元を使用した．なお本研究では生成の段階でテキストから音素継続長は予測せず，データセットから作成したコンテキストラベルを生成されたコンテキストラベルとして使用した．顔特徴量として，正規化された三次元特徴点の座標 4041 次元と正規化されたテクスチャの PCA 係数 128 次元，またそれらの一次，二次の動的特徴量を加えた計 12507 次元のベクトルを用いた．また DNN は表 1 に示したものを用いた．

3.2 提案法による顔動画像生成

実験結果を図 1 に示す．図は生成された動画とオリジナルの動画のあるフレームを比較したもので，左側が生成された動画，右側がオリジナルの動画である．口が開いているフレーム，閉じているフレーム共にオリジナルのような顔を再現することができた．一方でリップシンクの精度は高いとは言えず，今後検討の余地がある．また生成した顔を背景動画に貼り付けた際に，生成した顔とオリジナルの顔の境目が見えてしまっており不自然である．より自然な貼り付け方について今後検討していく必要がある．

(a) 口が開いているフレーム

(b) 口が閉じているフレーム

図 1: 生成動画 (左) とオリジナル (右) との比較

4 まとめ

本稿では Kinect で収録した三次元顔特徴点を用い，DNN に基づくテキストからの顔動画像生成手法を提案した．実際に生成した動画のフレームをオリジナルのものと比較することで，提案法によって顔動画像の生成が可能であることを示した．今後の課題として，3.2 節で触れたものに加え，学習データ数をより多くした場合や異なる DNN の構造を用いた場合について検討する必要がある．またいくつかの先行研究と比較して，生成動画の自然性や品質を主観評価する必要がある．

謝辞 本研究の一部は，JSPS 科研費 JP15H02720 の助成を得た．

参考文献

[1] Timothy F Cootes, Gareth J Edwards, and Christopher J Taylor. Active appearance models. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, No. 6, pp. 681–685, 2001.

[2] Mark JF Gales. Cluster adaptive training of hidden Markov models. *IEEE Transactions on Speech and Audio Processing*, Vol. 8, No. 4, pp. 417–428, 2000.

[3] Richard Anderson, Bjorn Stenger, Vincent Wan, and Roberto Cipolla. Expressive visual text-to-speech using active appearance models. In *Proc. Computer Vision and Pattern Recognitioan (CVPR)*, pp. 3382–3389, 2013.

[4] 佐藤一樹, 能勢隆, 伊藤彰則. “Animation Unit を用いた HMM･DNN によるテキストからのフォトリアリスティック顔動画像合成におけるカラー化の検討,” 信学技報, vol. 116, no. 220, pp. 67–72, 2016.

[5] Kinect for Windows SDK 2.0 Programming Guide : High definition face tracking. https://msdn.microsoft.com/en-us/library/dn785525.aspx.