ビデオ通話の動画を用いた対話システムの構築

Construction of a dialogue system using video of videotelephony

福井 空[1*]　稲葉 通将[1]
Sora Fukui[1]　Michimasa Inaba[1]

[1] 電気通信大学
[1] The University of Electro-Comunications

Abstract: This study aims to construct a dialogue system using a video of a real person as an interface. We use video of videotelephony divided into the questioner and the respondent and respond by dynamically switching the respondent's video according to the user's utterance. Since this system uses real persons' video, the user feels as if they are talking with the real person and is guaranteed to have spoken by the real person. We improve on the previous work and propose a new system that seamlessly switches between videos, creates a dialogue system from a single video of videotelephony, and provides nodding functions while the user speaks. The proposed system enables us to create a dialogue system with a high presence.

1　はじめに

本研究は, 実在の人物の動画をインターフェイスとして用いて対話システムを構築することを目的とする. 質問者と回答者が個別に録画された 2 人のビデオ通話動画を使用し, ユーザーの発話に応じて回答者の動画を動的に切り替えることで, ユーザーが回答者と仮想的に対話できるシステムを構築する.

この対話システムは実在する人物のビデオ通話の動画を用いるシステムであるため, 実際にその人物と話しているような臨場感が得られ, 発言が本人によるものであることも保証される.

既存の研究としてホロコースト生存者のインタビュー動画を用いて, 対話を行う対話システムが存在する [1]. 既存の研究では動画の切り替えにフェードイン・アウトを使用している点, ユーザーの発話中にうなずきや相槌がなく, 無反応である点, 動画を質問ごとに作成する必要があるため対話システム作成に時間がかかる点などが課題となっている.

本研究では動画を用いた既存の対話システムの問題点を改良し，より臨場感のある対話システムを作成することを目標とする.

1.1　提案システム

本研究で提案する対話システムの概要を図 1 に示す.

提案する対話システムは 3 つのモデルから構成される.

1 つ目は質問者と回答者のビデオ通話の動画を，回答者の回答動画と待機動画に分割するモデルである．このモデルを用いることにより，1 つの動画から対話システムの応答を作成することが可能となる.

2 つ目はユーザーの発話から応答として適切な動画を選択するモデルである．動画中の応答内容を人手で書き起こすのはコストが高いため，音声認識により自動で書き起こしを行う．また，ユーザの発話も音声認識結果を用いる，よって，入力とそれに対する応答の双方に音声認識誤りが含まれる可能性があることから，それらを考慮したロバストなモデルが必要である.

3 つ目は回答者の回答動画と待機動画をシームレスにつなぐような中間動画を生成するフレーム補完モデルである．本対話システムはユーザーの発話に回答する回答動画を再生し終えると，待機動画を再生する．回答動画と待機動画は連続した動画でないため，回答動画から待機動画へとつながるような中間動画を生成する必要がある.

本発表ではフレーム補完モデルに焦点をあて, 深層学習を用いて動画間のフレームを生成し, 動画の切れ目を意識させないような自然な遷移ができる手法を提案する.

2　提案手法

動画の遷移が自然に見えるようにするために, 前の動画 V_{pre} と次の動画 V_{next} の中間動画 V_{mid} を生成することが必要である.

*連絡先：電気通信大学情報理工学域経営社会情報プログラム
〒 182-8585 東京都調布市調布ヶ丘 1-5-1
E-mail: f1710528@uec.ac.jp

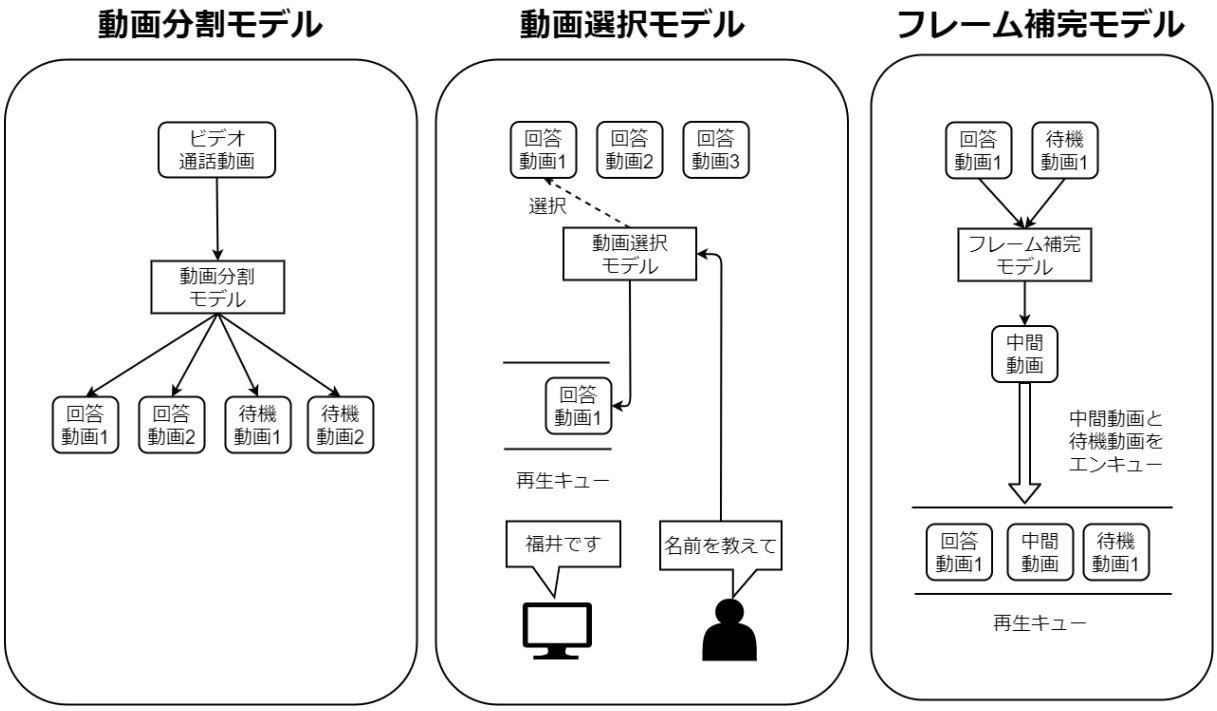

図 1: 提案する対話システムの概要

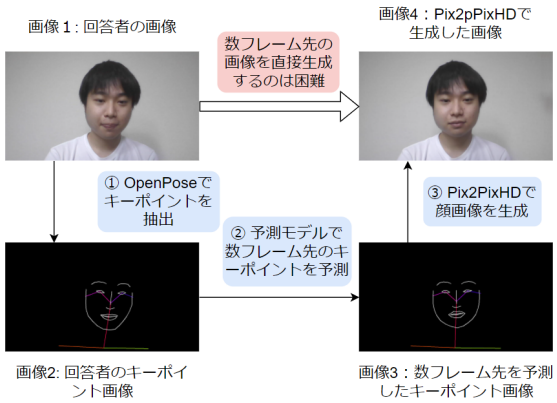

図 2: 提案手法

SuperSloMo[2] などの既存のフレーム補完技術は低 FPS の動画を高 FPS に変更するものであり, 今回のケースのように前の動画の最終フレームと次の動画の最初のフレームが大きく変化する場合にはうまく補完することができない.

2 画像の間を補完する動画を生成する研究もなされてはいるが [3], そのようなネットワークでも高画質な画像生成はできておらず, 入力に前の動画と次の動画を与えて高画質な中間動画を生成するようなモデルを作成することは困難であると考えられる. そこで提案手法では, まず動画に登場する回答者の体のキーポイント (座標情報) を推定し, 次に体の動きのみを予測し, 最後に体のキーポイントから回答者の画像を生成するという手法を提案する.

提案手法の概略を図 2 に示す. 本手法では OpenPose[4] を用いて動画に登場する回答者のキーポイントを推定する (sec. 2.1). 次に推定されたキーポイントを入力とし, フレーム予測モデルにより中間動画のキーポイントを予測する (sec. 2.2). 最後に, 画像から画像への変換を行う手法である pix2pixHD[5] を用い, 予測したキーポイント情報から中間の画像を生成する (sec. 2.3).

このアプローチにより元画像と同程度の高画質な画像生成するような補完を行うことが可能になる.

2.1 OpenPose によるキーポイントの抽出

OpenPose によるキーポイントの抽出は図 2 の①に該当する処理である. OpenPose は画像から人間の姿勢, 顔の特徴点を深層学習により推定することのできるライブラリであり, 姿勢や顔の特徴点の画像中の座標であるキーポイントを出力する. 本研究ではビデオ通話の動画を画像として切り出し, それぞれの画像に対して OpenPose を用いて図 2 の①のようにビデオ通話の動画に登場する回答者の顔, 姿勢のキーポイントを推定する.

2.2 フレームの予測

フレーム予測は図 2 の②に該当する処理である．フレーム予測は 2 通りの方法を用いて行う．

2.2.1 多層ニューラルネットによる補完

前の動画 V_{pre} の最後の n フレームと次の動画 V_{next} の最初の n フレームのキーポイント情報を入力として，中間フレームのキーポイント情報を出力するような，全結合の多層ニューラルネットワークを作成する．学習データにはビデオ通話の回答者の動画から推定したキーポイントを用いて，学習データのキーポイントと出力結果のキーポイントの平均二乗誤差を最小化するように学習する．このモデルにより前の動画 V_{pre} の最後のフレームと次の動画 V_{next} の最初のフレームの間を補完するような中間フレームのキーポイント情報を出力する．

2.2.2 線形補完

前の動画 V_{pre} の最後の 1 フレームと次の動画 V_{next} の最初の 1 フレームのキーポイント情報を用いて線形補完するモデルを作成する．前の動画 V_{pre} の最後の 1 フレームのキーポイントの座標の集合を P_{pre}，次の動画 V_{next} の最初の 1 フレームのキーポイントの座標の集合を P_{next} としたとき，中間フレームの 1 つのキーポイントの座標の集合 P_t は $t \in [0, 1]$ を用いて次の式で表される．

$$P_t = tP_{pre} + (1-t)P_{next} \tag{1}$$

また，中間フレームの生成枚数 m を求めるために，1 フレームあたりのすべてのキーポイントの移動距離の平均 d を次の式で計算する．なお，動画の全フレームの総数を L，1 フレームに含まれるキーポイントの総数を K，あるフレーム l におけるキーポイント k の座標ををを $p_{l,k}$ とする．

$$d = \frac{1}{K(L-1)} \sum_{l}^{L-1} \sum_{k}^{K} dist(p_{l,k}, p_{l+1,k}) \tag{2}$$

$dist(p_{l,k}, p_{l+1,k})$ は $p_{l,k}$ と $p_{l+1,k}$ のユークリッド距離を計算している．1 フレームあたりのすべてのキーポイントの移動距離の平均 d を用いて，中間フレームの生成枚数 m を次の式で計算する．

$$m = floor\left(\frac{\sum_k dist(p_{pre,k}, p_{next,k})}{Kd}\right) \tag{3}$$

$floor()$ は小数点以下の切り捨てである．

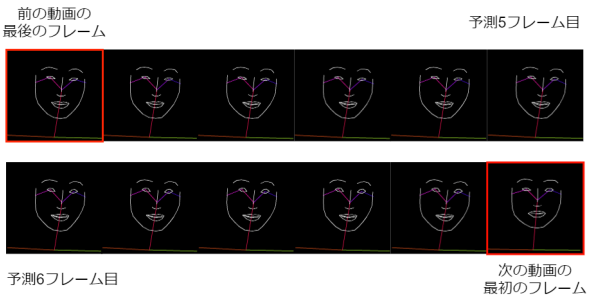

図 3: 多層ニューラルネットによるフレーム予測の出力結果

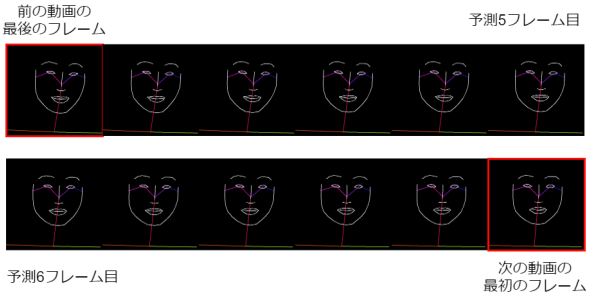

図 4: 線形補完によるフレーム予測の出力結果

2.3 Pix2PixHD による回答者の画像の生成

Pix2PixHD による回答者の画像の生成は 図 2 の③に該当する処理である．③の処理では，2.2 節で推定したキーポイントは単なる座標情報であることからキーポイントを画像化し，その画像から回答者の画像を生成する．Pix2PixHD の学習にはビデオ通話の動画から切り出した画像とその画像から OpenPose を用いて推定したキーポイントの画像のペアを用意し，キーポイントの画像から回答者の画像を生成するように学習させた．

3 実験

多層ニューラルネットと線形補完の 2 つのフレーム予測モデルが中間画像のキーポイントの予測を行えるかを確認するため実験を行った．

3.1 多層ニューラルネットによるフレーム予測モデル

入力データには，前の動画 V_{pre} の最後の 10 フレームと次の動画 V_{next} の最初の 10 フレームのキーポイント情報を用い，出力として中間の 10 フレームのキーポイントの予測を出力するモデルを学習させた．入力に用

いた動画は「あなたの名前はなんですか？」のような28個の質問に対し, 著者が回答したものである.

実験結果を図3に示す. 実験の結果, 予測されたキーポイントは前の動画の最後のフレームからほとんど動いておらず, 次の動画の最初のフレームと一致しなかった.

3.2 線形補完によるフレーム予測モデル

入力データには、 前の動画 V_{pre} の最後の1フレームと次の動画 V_{next} の最初の1フレームのキーポイント情報を用い, 式 (1), (2) に基づいて中間の m フレームのキーポイントを予測した．実験結果を図4に示す. 実験の結果, 予測されたキーポイントは前の動画の最終フレームと次の動画の最初のフレームがつながるように補完することができた. しかしながら, 予測されたキーポイントの動きは等速であるため，動画にすると若干の違和感があるように感じるものとなった.

4 まとめ

本研究では，単一のビデオ通話の動画から対話システムを作成することを目的とし, 構築のために必要な動画分割モデル, 動画選択モデル, フレーム補完モデルの3つのモデルを提案した. 特に本稿ではフレーム補間モデルに焦点を当て, OpenPose，Pix2PixHD を用いて, ビデオ通話の動画間の補完をするような手法を提案した. 本手法により, ビデオ通話の人物の動画を用いた臨場感のある対話システムが手軽に作成することが可能になると考えられる.

しかしながら, 現状のフレーム予測モデルは動画にすると違和感があるものとなってしまうという課題がある．多層ニューラルネットのフレーム予測モデルでは前後の動画をつなげるような中間フレームを予測することができないという課題が確認された．ネットワークのロスの設計や入力方法が適切でないことや, データ数が少ないことが原因であると考えられる．また線形補完モデルでは前後の動画をつなげるような中間フレームを予測することはできるものの，動きが等速で動画にした時の違和感が感じられた.

今後はフレーム予測モデルの改良に加え，動画分割モデルや動画選択モデルについての研究を進めていきたい.

参考文献

[1] David Traum, Andrew Jones, Kia Hays, Heather Maio, Oleg Alexander, Ron Artstein, Paul Debevec, Alesia Gainer, Kallirroi Georgila, Kathleen Haase, Karen Jungblut, Anton Leuski, Stephen Smith, and William Swartout. New Dimensions in Testimony: Digitally Preserving a Holocaust Survivor's Interactive Storytelling. In *Interactive Storytelling*, volume 9445, pages 269–281. Springer International Publishing, Copenhagen, Denmark, December 2015.

[2] Huaizu Jiang, Deqing Sun, Varun Jampani, Ming-Hsuan Yang, Erik Learned-Miller, and Jan Kautz. Super slomo: High quality estimation of multiple intermediate frames for video interpolation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[3] Tsun-Hsuan Wang, Yen-Chi Cheng, Chieh Hubert Lin, Hwann-Tzong Chen, and Min Sun. Point-to-point video generation. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[4] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In *CVPR*, 2016.

[5] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.