J-027

特定人物の身振りをアバターで再現するシステム

An Avatar Generation System Reproducing Gesture of a Specific Person

田原 俊一† 服部 元†
Shunichi Tahara Gen Hattori

1. はじめに

近年，対話エージェントは広く普及しており，特定人物の姿を表現した CG アバター（「特定人物アバター」と定義）も登場している．特定人物アバターと対話をすることで，当該人物と対話をした気になり，ユーザの満足感が向上すると考えられる．人間の個性は仕草の中に現れやすい[1]ことから，特定人物アバターを実現するための要素として特に身振りは重要視される．現状では，アバターの発話から標準的な視覚表現を自動生成するシステム[2]が存在するが，特定人物に特徴的な視覚表現を自動生成するシステムは存在しない．本稿では，特定人物が映る複数の映像から，当該人物に特徴的な身振りの映像を抽出して，身振りを特定人物アバターで再現するシステムを提案し，有効性を評価する．

2. 関連研究

石井ら[2]は，様々な人同士での対話を映像収録して，発話や視覚表現（表情，身振りなど）を視覚表現生成器に学習することで，アバターの発話内容に応じて，視覚表現を自動生成するシステムを提案している．しかし，生成した視覚表現は標準的な視覚表現であり，特定人物の姿を再現したアバターにこのシステムを適用すると，当該人物らしい表現ができず，視覚表現に大きな違和感が生じてしまう課題がある．

3. 提案システム

3.1 提案システムの概要

本稿では，特定人物とそれ以外の人物（「非特定人物」と定義）が映る映像から，特定人物に特徴的な身振りを解析し，特定人物アバターの発話内容に合わせて、当該人物らしい身振りを再現するシステムを提案する．図 1 に提案システムの概要を示す．

始めに，特定人物が映る映像から，当該人物が発言している部分の映像をセンテンス毎に人手で抽出する．同様にして，非特定人物が映る映像から，任意の人物が発言している部分の映像をセンテンス毎に人手で抽出する．それぞれ抽出した映像を特定人物映像群，非特定人物映像群と定義する．

次に，以下の 3 ステップで特定人物の身振りをアバターで再現する．(1)発話種類推定機能は，特定人物映像群と非特定人物映像群それぞれにおいて，抽出した各映像の発言がどのような種類の発話であるか推定する．(2)身振り統計解析機能では，発話の種類毎に映像をクラスタリングする．特定人物映像群と非特定人物映像群のクラスタリング結果を比較した上で，前者に特徴的に出現するクラスタを決定する．(3)身振り選択機能では，(2)で決定したクラスタ内に存在する映像の中から代表的な映像を 1 つ選択し，発話種類毎に映像 DB に保存する．

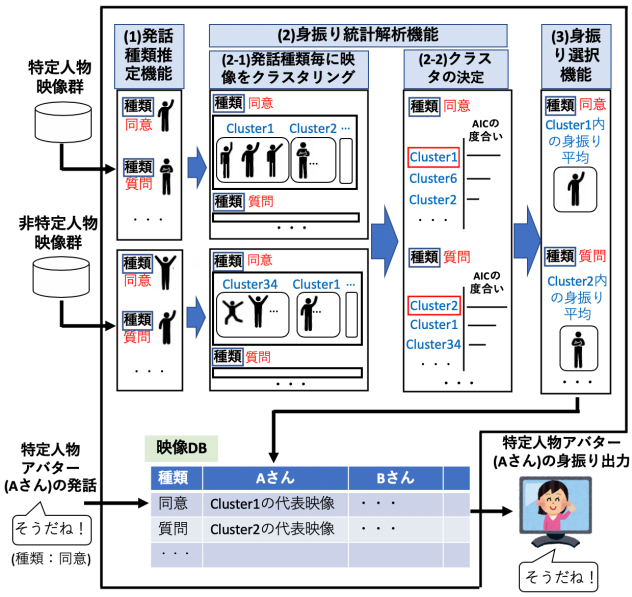

図 1：提案システムの概要

最後に，特定人物アバターの発話をシステムに入力すると，発話種類の推定を行い，映像 DB に記録されている発話種類や特定人物の名前に応じて，特定人物アバターの身振りが画面上に出力される．

3.2 発話種類推定機能

特定人物アバターの発話内容に応じた身振りを出力するため，特定人物映像群，非特定人物映像群における各映像中の発話の種類を COTOHA API ¶を用いて推定し，発話種類毎に映像を分類する．具体的な発話種類としては，情報提供 3 種（Positive, Negative, Neutral）および，同意，質問，その他の 6 種類とする．

3.3 身振り統計解析機能

特定人物に特徴的な身振り映像を取得するため，3.2 節で推定した発話の種類毎に映像をクラスタリングし，特定人物映像群において特徴的に出現するクラスタを決定する．

クラスタの決定は次の手順で行う．(2-1)特定人物映像群及び非特定人物映像群における各映像は，複数の静止画から構成される．文献[3]の手法を用いて，各静止画に映っている人体の骨格点座標を推定する．発話種類毎に，各映像における全静止画の骨格点座標の平均から最も距離が近い

† KDDI 総合研究所, KDDI Research, Inc.

¶ https://api.ce-cotoha.com/contents/index.html

静止画（「代表静止画）と定義）を 1 つ決定し，代表静止画の骨格点座標を特徴量に変換する．次に，k-means を用いて，各映像における代表静止画のクラスタリングを行う。その結果を映像のクラスタリング結果に反映する．(2-2)特定人物映像群と非特定人物映像群における映像のクラスタリング結果を比較し，特定人物映像群における各クラスタ内の映像の数が高頻度に偏って出現する度合いを AIC（赤池情報量基準）[4]を用いて算出する．

3.4 身振り選択機能

特定人物アバターの身振りを決定するため，3.3 節で算出した AIC の度合いが上位のクラスタに含まれる各映像の代表静止画の骨格点座標の平均を算出し、平均に最も距離が近い代表静止画を選択する．代表静止画が属する映像を映像 DB に記録する．

4. 提案システムの評価

4.1 実験概要

提案システムで出力した特定人物アバターの身振りが，当該人物らしい身振りか評価をするために，標準的な身振りを出力するシステム[2]（「既存システム」と定義）と比較し，主観評価実験を行なった．

まず，特定人物が映っている映像を実験参加者に見せて，特定人物の身振りを把握させる．次に，図 2 のように骨格姿のアバター（「骨格アバター」と定義）と話者がテキストで対話をしている複数の映像を実験参加者に見せる．骨格アバターの発話は人手で入力を行う．映像視聴後に「骨格アバターの身振りが特定人物らしいと思ったか」と問い，映像毎に，-3（全く思わない）から 3（非常に思う）までの評価値と，その評価値をつけた理由をアンケートに回答させる．5 名が本実験に参加した．

収録する対話について述べる．既存及び提案システムで身振りを出力する場合それぞれにおいて，指定した 4 つの話題に基づく対話を行う．1 対話は 8 分間であり，2 手法それぞれについて合計 32 分間の対話映像を実験参加者に見せる．各話題の対話は，両システムを用いた場合で同一の対話を用いる．

既存システムで用いる映像，及び提案システムの非特定人物映像群における映像は， YouTube から取得した動画を使用した．提案システムの特定人物映像群における映像は，特定人物が対話をしている映像を使用した．尚，提案システムで用いる k-means のクラスタ数は，チューニングを行った上で 10 に設定した．

4.2 評価結果

実験結果を表 1 に示す．特定人物らしさの平均の有意差を Mann-Whitney の U 検定により分析した．有意水準 1(%)で，既存システムと提案システムの間で有意差が認められた（Z = 4.687）．これらの結果から，実験参加者は，既存システムよりも提案システムで選択された身振りに特定人物らしさを感じたことが示された． 骨格姿のアバターの身振り評価という，実験参加者にとって回答の判断が困難な実験において，提案システムの特定人物らしさの平均が 1.50 となり，7 段階中 5 段階以上であることから提案システムによる効果は大きいと言える．

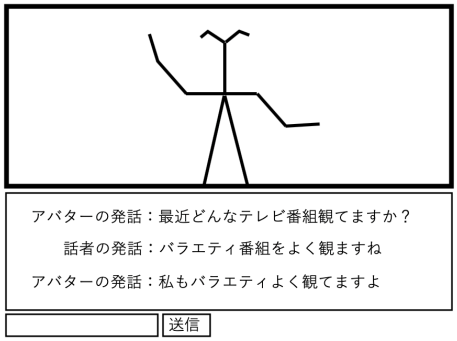

図 2：実験参加者に見せる対話映像の例

表 1：既存/提案システムにおける特定人物らしさの平均

既存	提案
-0.65	1.50

既存システムでは，「特定人物には見られない動作が多くあったように感じた」，「手振りが大きく特定人物のイメージと違う」といったネガティブなコメントが多かった．

提案システムでは「特定人物の身振りの特徴をよく掴んでいるように見える」，「髪を触る動きが似ている気がする」，「両手で円を作るようにしている動きが似ている」等の，特定人物らしい身振りが実現できているといったポジティブなコメントが多かった．特定人物が映る映像を確認すると，手で髪を触る身振りや，両手を使って説明をする仕草が多く出現しており，特定人物の特徴的な身振りが骨格アバターの身振りに反映されていたと考えられる．

一方，提案システムでは「特定人物らしい動作が見られたが、身振りがパターン化されているように感じた」といったネガティブなコメントも見られた．骨格アバターの身振りのバリエーションを増やすことが今後の課題であると考え，身振りの自然さの向上を目指す．

5. まとめと今後の課題

本稿では，特定人物が映る映像から当該人物に特徴的な身振りを解析し，特定人物アバターで身振りを再現するシステムを提案した．実験によって，既存システムと比較した際，提案システムで出力した身振りの特定人物らしさが有意に高くなることが示された．骨格アバターの身振りの自然さを向上すべく，今後の課題としてアバターの身振りのバリエーションを増やすことが挙げられる．

参考文献

[1] 岡隆一, 西村拓一, 向井理朗, "しぐさで伝える." 電子情報通信学会誌 82.4, 332-339(1999).
[2] 石井亮, 片山太一, 東中竜一郎, 富田準二, "発話言語に基づく身体モーションの自動生成", 分散協調とモバイルシンポジウム 2018 論文集, 1863-1868 (2018).
[3] XU, Jianfeng., TASAKA, Kazuyuki., YANAGIHARA, Hiromasa., “Beyond Two-stream: Skeleton-based Three-stream Networks for Action Recognition in Videos.” In 2018 24th International Conference on Pattern Recognition. 1567-1573 (2018).
[4] AKAIKE, Hirotugu. “A new look at the statistical model identification”, IEEE transactions on automatic control 19.6, 716-723(1974).