1ZE-06

ロボットの経験を伝える発話とジェスチャの同時生成システム

下山 香音† 大塚 洋平‡ 金重 有哉‡ 木本 充彦† 今井 倫太†
慶應義塾大学理工学部† 慶應義塾大学大学院理工学研究科‡

1. はじめに

本研究では，ロボットが経験した情景を伝えるための発話とジェスチャの同時生成システムを提案する．人間はコミュニケーションを行う際に，マルチモーダルな振る舞いを通して，発話だけの場合よりもより多くの情報を伝達している．同様に，ロボットもマルチモーダルな振る舞いにより人に与える情報が増し，発話内容がイメージしやすくなるだろう．

[1]では，ジェスチャは単体で存在せず，発話表現を補足するものとしている．これまで，Kucherenko *et al.*は音声を入力とし音声に合わせた拍子のジェスチャの生成を行なった [2]．また，Nihei *et al.*は画像を入力とし画像中の物体の形を表現するジェスチャの生成を行なった [3]．しかし，[2]では音声の時間による変化が考慮されているが，ジェスチャで発話内容を補足できない．また，[3]では発話内容の補足が行えるが，物体の移動など，時間による変化を表現できない．そのため，ロボットの発話とジェスチャを見た人間がロボットの想定外の解釈を行い，誤解が生じる可能性がある．

そこで，本研究では動画を入力として iconic ジェスチャと発話の生成を行うシステム，VISG（Video-based Iconic gesture and Speech Generator）を提案する．VISG で生成されたジェスチャは発話内容を補足することができ，かつ，時間による変化を表現することができる．

2. 予備実験：人の説明方法収集

2.1 概要

VISG では，動画中の言及対象の決定と発話の生成をルールベースで行う．そこで，ルールの設定のために日本人が動画の説明においてどのような表現を行うのか調査した．

実験協力者には 10 秒程度の動画を数本観てもらい，その動画を観たことがない人に対して説明する際にどのように話すのか一文で回答してもらった．これを 2 本の動画に対して 20 人，10 本の動画に対して 30 人に実施した．

動画は動画キャプショニングデータセット VaTeX [4]からランダムに選出した．なお，極端に画質が悪い動画と途中でテロップが表示される動画については除外した．

2.2 結果

動画と収集した説明文を分析した結果，3 つの傾向がわかった．まず，説明文の主旨となる要素は，「ものが落ちた」や「止まった」といった動画の始めの状態から変化があるものが多く，動画中に状態の変化がない場合は「踊っている」や「移動している」といった，動きのあるものが多かった．また，説明文の主旨とは関係の薄い要素についても言及されることがあった．これは，その説明を受けた人がより情景を思い浮かべやすくするためだと考えられる．

これらをもとにして考案した，言及対象と発話の生成に用いる定義を表 1 にまとめる．

表 1 分析結果を用いた定義

表現対象	定義
イベント	始めの状態から変化がある要素
前景	イベント，あるいはイベントではないものの動いている要素 説明文の主旨となる
背景	説明文の主旨ではない要素

3. 提案システム

[5]はジェスチャを 4 つに分類しており，それぞれ，形や動作を表現する iconic，抽象的な概念を表現する metaphoric，指差しの deictic，発話に併せた拍子の beat である．

そこで本研究では，動画を入力として iconic ジェスチャと発話の生成を行うシステム VISG を提案する．

図 1 のように，VISG は動画から物体とその移動を抽出し言及対象を決定し，それをもとにジェスチャと発話を生成する．システムの構成を図 2 に示す．

Simultaneous generation of a robot's speech and gesture to convey an experience
† Faculty of Science and Technology, Keio University
‡ Graduate School of Science and Technology, Keio University

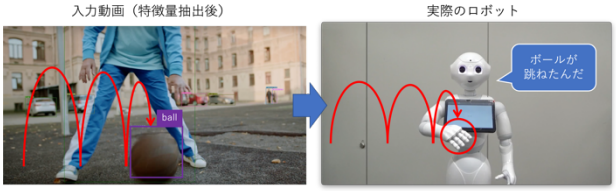

図 1 ジェスチャ生成例

図 2 システムの構成

3.1 動画特徴量抽出モジュール

動画の物体検出および物体追跡のために [6]を用いる．これにより，物体のクラスと各フレームにおける座標を取得する．

3.2 言及対象決定

2 章で得た結果をもとに，ルールベースで言及対象を決定する．背景は，前景とは異なるクラスで認識された物体を選ぶ．

動画特徴量抽出モジュールで得た結果を入力とし，検出された物体の座標を使用して用意した基準にどれほど当てはまるか計算を行う．用意した基準は，表 2 の通りである．

表 2 言及対象の決定に用いる基準

前景	イベント	落ちる
		動きが止まる
		動きだす
	イベントではない	移動が速い
		移動距離が長い
		画面外側から内側へ移動する
背景		移動が遅い
		同じクラスの物体の数が多い
		物体が大きい

3.3 ジェスチャ生成

言及対象の動画中の座標をもとに，ロボットを正面から見た時のロボットの手の位置が言及対象の座標と対応するようにジェスチャを生成する．

3.4 発話生成

言及対象の決定に用いられた基準と物体のクラスをもとに，ルールベースで発話の生成を行う．

3.5 ロボット

ロボットは腕の自由度が人間と近く表現の幅が広い Pepper を使用した．

4. まとめ

動画を入力として iconic ジェスチャと発話の生成を行うシステム VISG を提案した．今後，実際にロボット視点の映像を使用してジェスチャと発話の生成を行い，その効果を測る．

謝辞

本研究の一部は JST，CREST，JPMJCR19A1，および科研費 JP20K19897 の助成を受けたものです．

参考文献

[1] A. Kendon, Gesture : Visible Action as Utterance, Cambridge University Press, 2004.

[2] T. Kucherenko, D. Hasegawa, G. E. Henter, N. Kaneko , H. Kjellström, “Analyzing Input and Output Representations for Speech-Driven Gesture Generation, ” roceedings of the 19th ACM International Conference on Intelligent Virtual Agents, 2019.

[3] F. Nihei, Y. Nakano, R. Higashinaka , R. Ishii, “ Determining Iconic Gesture Forms Based on Entity Image Representation, ” 2019 International Conference on Multimodal Interaction, 2019.

[4] X. Wang, J. Wu, J. Chen, L. Li, Y.-F. Wang , W. Y. Wang, “VaTeX: A Large-Scale, High-Quality Multilingual Dataset for Video-and-Language Research,” The IEEE International Conference on Computer Vision (ICCV), 2019.

[5] D. McNeill, Hand and mind : what gestures reveal about thought, University of Chicago Press, 1992.

[6] Mike. Available: https://github.com/mikel-brostrom/Yolov5_DeepSort_Pytorch.