動画からうなずきを抽出する試み

Attempt to Extract Nods from Videos

伊藤　敏[*1]，　井上祥史[*2]，　鷲野 嘉映[*3]
Satoshi ITOU[*1], Shoshi INOUE[*2], Kaei WASHINO[*3]
[*1] 岐阜聖徳学園大学
[*1]Gifu Shotoku Gakuen University
[*2] 岩手大学
[*2]Iwate University
[*3] 愛知みずほ短期大学
[*3] Aichi Mizuho Junior College

Email: itous@gifu.shotoku.ac.jp

あらまし：会話解析をするためのツールとして，動画から顔検出をし，顔の特徴点の軌跡情報から多層パーセプトロンを用いてモデルを作成し、会話解析に有用と思われる，うなずきなどを推定する方法を試みた．また，カメラ 1 台で複数者の会話中の正面動画を取得する方法を提案した．その結果，会話中の聞き手の行動を推定し，目視による結果と一致することを示した．
キーワード：**会話**，録画，顔検出，うなずき

1.　はじめに

会話などにおける聞き手の言語活動以外の行動(ノンバーバルコミュニケーション)が，話者にとって「話しやすさ」などに関係すると思われる(1)．会話時のノンバーバルコミュニケーションのうち「うなずき」の検出をする試みとして会話者の頭部に慣性センサを取り付け，動きを検出する試みや(2)，カメラを用いて動画から検出する試みなどがある(3),(4)．また，我々も会話者の頭部にセンサを装着して，行動を観測する試みを行ってきた(5)．

ビデオカメラによる観測は，会話者に装置を装着せず非接触で行動を観測可能あるため，会話者への負荷が少ない．一方で，記録された動画を目視により解析した場合，解析者に大きな負担がかかる．これらを機械学習手法により分類する試みがある(6)．しかし、うなずきなどの行動には多様性があり、汎用的に利用可能な自動化は難しい．

本稿の目的は，会話行動を記録した動画から「特定のうなずき」動作を抽出する事である．ここで「特定」とは解析対象者のうなずき特徴に沿って学習モデルを作成し推定することを指す．これにより解析対象のうなずき解析に資するものと考える．

本稿では，2 章で動画からのデータ取得の方法と顔動作推定の方法を記述し，3 章で応用例として，二者対話での会話を録画し，行動の抽出を試み，4 章でまとめる．

2.　方法

顔の動きを分類するために、顔の動作の軌跡を記録し、機械学習を用いて分類するモデルを作成した．そのモデルを用いて、顔の動きを推定した．

2.1 顔座標の抽出

顔検出と顔の特徴点を検出するのに、機械学習ライブラリである dlib(7)、または mediapipe(8)を用いた．得られた顔器官の座標のうち鼻頭、両目尻の 3 座標点を記録し解析に供した．これら 3 点を選んだのは、表情変化などによる相対的座標変化が小さいため、また個人差が出にくいと判断できるためである(9)．

2.2 顔座標の軌跡から動作推定

分類推定する「動作」を「何もしない(Normal)」、「うなずき(Nod)」と「否定(Disagree)」の 3 種とした．顔の動き動作を学習するために、USB カメラを用いて、上記 3 動作中の、鼻頭、両目尻の 3 点の軌跡を記録した．これらの 3 動作に 3 特徴点の軌跡を深層学習し、分類を試みた．

図 1 に，mediapipe を用いた Landmarks（赤色）と検出対象とした顔の特徴点 3 点（青点）を示す．これらの座標は 30fps で数値として保存され，行動の分類に用いられた．「うなずき」は顔の短時間での上下運動、否定は左右動とみなし，顔の中心部に位置する鼻頭、両目尻の軌跡変動から推測する．

モデル作成には、うなずき、否定行動は 1 秒程度の短時間で完結するため、過去 16 枚（0.53 秒）を 1 かたまりとして解析した．なお、動画収録環境による差をなくすため、次の規格化を行った．1) 入力数

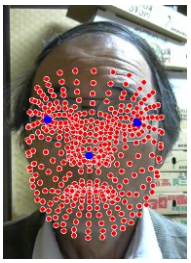

図 1．顔検出と特徴点の座標 mediapipe の場合

値は得られた座標を動画の解像度で割り、2) 各かたまりの最初の鼻頭，両目尻からの相対位置を用いた(かたまりのデータ数=16 枚×3 点×2 点(x,y) - 6=90).

軌跡から動作を推定するモデル作成には、入力層に 90、隠れ層を 2 層、出力層を 3 (推定する動作数) とした多層パーセプトロンを用いた．各動作を 30 から 180 秒程度のデータで学習を行い、モデルを構築した．

構築したモデルを用いて、USB カメラや記録動画から抽出された鼻頭、両目尻の座標データを取得し、先のモデルを用いて推定をし、各時間での推定動作を記録した．

2.3 正面動画での動作推定

学習したモデルの検証のために正面から撮影された動画で推定を行った．学習モデル構築に用いた人物とは別人による、3 動作を 15 秒から 30 秒程度、繰り返しうなずきの速さや大きさそして頻度を変えて行った行った動画から推定した結果を表 1 に示す．いずれの場合も 80%を超える推定率である．これはモデルが有効に構築され、また鼻頭、両目尻の軌跡を用いることで個人差がほとんどなくなったためと推察される．

3. 会話中の行動分析への適応

実際の会話中の行動推定を行った．

3.1 360 度カメラでの動作推定

実際の会話で、正面動画取得は難しい．そこで、正面動画が取得可能な全天球 360 度カメラで録画して推定した. 360 度カメラ映像は mediapipe では顔検出ができない．そこで dlib で顔検出をし、鼻頭両目尻の座標を取得した．そのデータを用いて動作推定した結果を図 2 に示す．目視でラベリングした結果と推定結果が比較的良く一致している．なお、ラベリングには ELAN を用いた[10].

表 1. 正面動画からの動作推定結果

	person1 推定結果			person2 推定結果		
実際の動作	Normal	Nod	Disagree	Normal	Nod	Disagree
Normal	0.946	0	0.006	0.863	0.011	0.021
Nod	0.054	1.000	0.071	0.134	0.974	0.002
Disagree	0	0	0.923	0.002	0.016	0.977

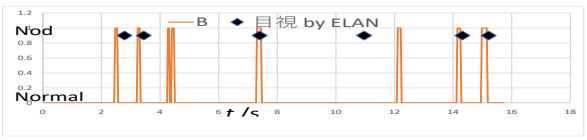

図 2. 360 度カメラ動画からの Nod 推定

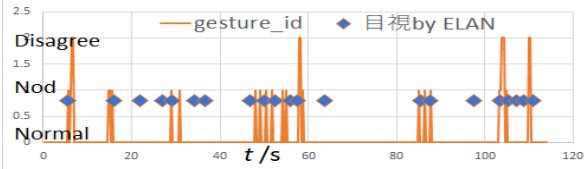

図 3. 斜め動画からの Nod 推定

3.2 斜め動画での動作推定

過去に収録された会話動画や YouTube などは、会話者と聞き手が斜めに写った場合が多い. その場合、正面動画から構築したモデルでは推定精度が下がる. そこで斜め動画から構築したモデルを用いた結果を図 3 に示す．縦軸 1 が Nod、2 が Disagree を表す．動画では挨拶と Nod 以外の動作はない．会話開始時と終了時に挨拶が行われ、それらが Disagree として推定されているが、それらを除けば、概ね Nod 動作を推定できていると思われる．

また少ないデータで学習モデルが構築可能なため（3 動作で 2，3 分の学習時間）利便性があると思われる．

4. まとめ

対話の場合、聞き手のうなずき動作は多様であり、多様なうなずき動作に対応したモデルを構築することは困難である．本方法を用いることで、撮影された角度やうなずきなどの動作の「癖」を事前に把握し、学習モデルを構築することで動作推定ができると考えられる．これらの結果より，本方法は会話分析のツールとして利用可能であると考える．

うなずきに関して大塚容子氏から助言を受けた．本研究の一部は科研費（19K03178, 20K03164）の助成を受けた．

参考文献

(1) 泉子・K・メイナード:“会話分析”，くろしお出版，東京 (1993)
(2) 斎賀弘泰，角康之，西田豊明:“多人数会話におけるうなずきの会話制御としての機能分析”，情報処理学会研究報告, Vol.2010-UBI-26 No.1, pp1-8（2010）
(3) Morency, L., de Kok, I. and Gratch, J.: "Context-based recognition during human interactions: automatic feature selection and encoding dictionary", Proceedings of the 10th international conference on Multimodal interfaces, ACM New York, NY, USA, pp.181-188 (2008).
(4) 伊藤敏、大塚容子、鷲野嘉映:“動画から顔の動きを抽出する試み”，教育システム情報学会第 43 回全国大会, pp401-4023（2018）
(5) 伊藤敏，王琳琳，鷲野嘉映，井上祥史:“慣性センサを用いた行動検出試行”，教育システム情報学会 2016 年度第 2 回研究会, pp9-13（2016）
(6) 曽根田悠介，中村優吾，松田裕貴，荒川豊，安本 慶一:“ミーティング映像からの発話およびマイクロ動作識別手法”，情報処理学会研究報告，2020-UBI-065, pp-1-8 (2020).
(7) http://dlib.net/ 2022 年 5 月 15 日確認
(8) https://google.github.io/mediapipe/solutions/face_mesh 2022 年 5 月 15 日確認
(9) 日本人約 200 名の頭部寸法 外眼角幅 Biectocanthion breadth 河内まき子・持丸正明, 2008 : 日本人頭部寸法データベース 2001，産業技術総合研究所 H16PRO-212.
(10) ELAN (Version 6.3) [Computer software]. (2022). Nijmegen: Max Planck Institute for Psycholinguistics. Retrieved from https://archive.mpi.nl/tla/elan