一般社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report
CNR2025-4,HCS2025-46(2025-10)

マルチモーダル情報に基づくオンライン会議参加者の同意・不同意の推定

齊藤 寛己† 佐藤 矯汰郎† 大和 淳司††

†工学院大学大学院 工学研究科情報学専攻 〒192-0015 東京都八王子市中野町 2665-1
††工学院大学 情報学部情報科学科 〒192-0015 東京都八王子市中野町 2665-1
E-mail: †{em24024, em24025}@ns.kogakuin.ac.jp, ††yamato@cc.kogakuin.ac.jp

あらまし 人の対話において，非言語情報は相手の感情や会話の雰囲気を把握するうえで重要である．しかし現在のオンライン会議システムでは対面会議に比べて非言語情報が伝達されにくく，会議参加者の意図や感情の把握が困難である．本研究ではオンライン会議中の映像から参加者の同意・不同意を推定し，可視化することで会議の状況把握を支援することを目的とする．Action Unit および頭部運動の特徴量を用いて SVM によるマルチモーダル推定器を構築し，精度検証を行った．先行研究より被験者の同意時，不同意時の表情表出には大きな個人差が存在する．そのため，個人差を考慮しない汎用モデルでは認識精度が低くなるという課題がある．被験者間の平均識別率を距離尺度としたクラスタリングにより抽出した類似グループと全被験者を含むグループで精度比較を行った結果，類似グループの不同意ラベルの推定精度が相対的に高い値を示した．また特徴量の有効性を検証した結果，時間的構造を考慮していなかったために頭部運動の特徴量には有効性が示されなかった．

キーワード オンライン会議，同意，不同意，マルチモーダル情報

Estimating agreement and disagreement among online meeting participants based on multimodal information

Hiroki SAITO† Kyotaro SATO† and Junji YAMATO††

†Informatics Program, Graduate School of Engineering, Kogakuin University 2665-1 Nakano-machi, Tokyo, 192-0015 Japan
††Department of Information Science, School of Informatics, Kogakuin University 2665-1 Nakano-machi, Tokyo, 192-0015 Japan
E-mail: †{em24024, em24025}@ns.kogakuin.ac.jp, ††yamato@cc.kogakuin.ac.jp

Abstract Nonverbal cues are essential for interpreting emotions and intentions in dialogue, yet online meeting platforms hinder their transmission compared to face-to-face interactions. This study aims to enhance situational awareness in online meetings by estimating and visualizing participants' agreement or disagreement from video data. We developed a multimodal classifier using an RBF-kernel SVM with Action Unit (AU) and head-movement features, and evaluated its performance. A key challenge is that inter-individual variability in facial expressions reduces the accuracy of general-purpose models. To address this, we introduced a subgroup approach in which participants were clustered using a distance metric defined by pairwisecross-subject averaged recognition rates, and subgroups were then formed through hierarchical clustering. We compared this method with an all-subject model and found that subgroup-based models achieved higher classification accuracy. Feature analysis further indicated that head-movement cues contributed little, likely due to the absence of temporal modeling.

Keywords Online meeting, Agreement, Disagreement, Multimodal information

1. はじめに

近年，新型コロナウイルス感染症（COVID-19）の感染拡大を機にオンライン会議システムの普及が進み，社会全体で利用者が増加した[1]．オンライン会議システムには Zoom や Google Meet，Microsoft Teams など多様なサービスが存在し，利用環境や用途に応じて選択される．人の対話では言語情報に加え，表情や頭部運動といった非言語情報も意思疎通に重要である

[2][3]. 特に非言語情報は相手の感情や会話の雰囲気を把握するうえで重要な役割を担う．しかし現在のオンライン会議は対面会議に比べて非言語情報が伝達されにくく,会議参加者の意図や感情の把握が困難である．

オンライン会議における非言語情報を利用した研究は多く行われているが，非言語情報の種類により様々な利害得失があることが報告されている．表情を利用した研究では，Kruzic ら[4]は自動抽出された表情特徴量が態度予測に資する可能性を示す一方で，カメラとの距離や顔への手の接触，顔を背ける動作によるトラッキングエラー，マイクロ表情の検出が困難であると指摘している．頭部運動を利用した研究では，Kang ら[5]は低コストで取得可能であり，オンライン環境で失われがちなアイコンタクトの代替として活用できると述べている．また Levordashka ら[6]は同じ頭部運動でも文脈や場面によって解釈が変わり得るため留意が必要と示している．

本研究は，オンライン会議中の映像から，会議参加者の同意・不同意の心情を推定し，可視化することによってオンライン会議の状況把握を支援することを目的とする．本稿ではマルチモーダルな推定器を構築するために特徴量に表情,頭部運動のデータを使用する．表情特徴量には OpenFace 2.0[7]により抽出した 17 個の Action Unit[8]（以下，AU）の強度データを用いる．AU とは顔の解剖学の知見をもとに分類された顔の個々の筋肉または筋肉群の基本的な行動単位である．OpenFace 2.0 で検出可能な 17 個の AU の番号とその対応部位，動作を表 1 に示す．頭部運動特徴量には表情特徴量と同様に OpenFace 2.0 を用いて抽出されたラジアン単位の X，Y，Z 各軸の回転角（pitch，yaw，roll）を使用する．頭部回転角の定義を図 1 に示す．また OpenFace 2.0 による認識例を図 2 に示す．OpenFace 2.0 における解析時のウィンドウには,左から順に顔領域・ランドマーク検出結果，HOG 特徴の抽出結果，頭部姿勢・視線方向，各 AU の推定値が表示される．

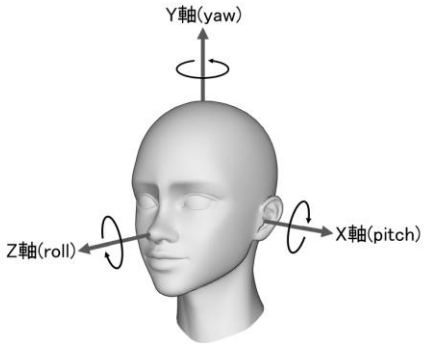

図 1　頭部回転角

表 1　OpenFace2.0 で検出可能な AU

AU	部位・動作
1	眉の内側を上げる
2	眉の外側を上げる
4	眉を下げる
5	上瞼を上げる
6	頬を持ち上げる
7	瞼を緊張させる
9	鼻に皺を寄せる
10	上唇を上げる
12	唇両端を引き上げる
14	えくぼを作る
15	唇両端を下げる
17	顎を上げる
20	唇両端を横に引く
23	唇を固く閉じる
25	顎を下げずに唇を開く
26	顎を下げて唇を開く
45	瞬く

図 2　OpenFace 2.0 による認識例

2. 関連研究

三浦ら[9]はマルチ対話コーパス（Hazumi）のオンライン収録データを用いたユーザーの感情推定モデルの開発を目指した．マルチモーダル対話システムの開発における課題として「マルチモーダルデータセットの収集コストの高さ」,「非言語情報の表出における個人差に起因するノイズ」を挙げている．これらの課題解決のため，少量データ，少ない学習ステップで分類タスクに適応可能なメタ学習手法 MAML（Model-Agnostic Meta-learning）を応用し，有効性の検証を行った．その結果，個人差を前提とした学習，評価の設計とメタ学習による個人適用がユーザーの感情推定の性能向上に有効であることを示した．

また先行研究として齊藤ら[10]はオンライン会議から得られた顔画像を用い，同意・不同意の推定に有効な特徴量を検討した．会議参加者の個人特徴を分析するうえで，同一の表情変化であっても同意時に強く表

情表出する参加者と不同意時に強く表出する参加者が混在し，構築するモデルは表情表出の個人差の影響を大きく受けることを明らかにした．このため，推定に有効な特徴量を選定には，参加者の非言語情報表出に基づくクラスタリングとクラスタ別評価が必要となることが示唆された．

以上より，精度向上には個人差の考慮が求められる．本稿では類似度が高い人をクラスタリングすることで個人差の影響を減らし，推定精度が向上するか分析する．また表情および頭部運動単体の特徴量を用いる推定器と表情＋頭部運動のマルチモーダルな推定器の精度を比較することで使用特徴量による精度の違いを確認する．

3. データセット

本研究では，クラウドワークスを利用し収集した 21 名（A～U，男性 15 名，女性 6 名，10 代～60 代）のオンライン会議映像データをデータセットとして使用した．クラウドワークスによる映像データ収集のプロセスは予備実験と本実験の 2 段階に分けて実施した．予備実験はアノテーションツールである ELAN[11]の操作確認を目的とし，予備実験の合格者のみが本実験に参加した．本実験はオンライン会議の映像データ収集および映像に対応した同意・不同意のラベル付与を目的とし，Zoom による 4 人 1 組のオンライン会議を実施した．会議の議題は「若者の選挙の投票率を向上させるための案の検討」とし，参加者が同意・不同意の意見を持ちやすい議題設定とした．会議は自身の案の説明と質疑応答を行う形式で進行した．会議参加者同士の関係は実験予定日に都合の合う人で 4 人組を組んだため，年齢や性別の割合などは考慮しておらず，初対面である．オンライン会議の録画は Zoom 内の録画機能を利用して行った．録画された映像データの解像度は 1280×720px であり，フレームレートは 25fps である．実施したオンライン会議の様子を図 3 に示す．OpenFace 2.0 により映像データから 1 フレームごとの 17 個の AU の値および 3 個の頭部運動の値を抽出した．

本実験であるオンライン会議終了後，各参加者は自身の録画映像に対してアノテーションを実施し，自身の発話区間の除く同意・不同意と感じた全ての区間に同意または不同意のラベル付けを行った．

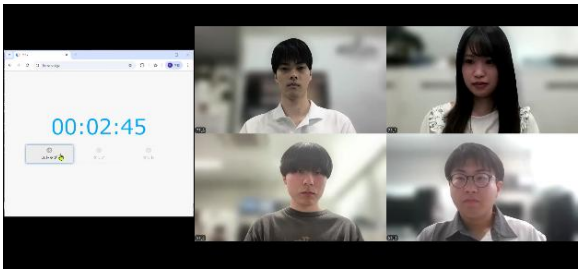

図 3　実施したオンライン会議の様子

4. 実験

4.1. クラスタリング

クラスタリングのための被験者間の距離を以下の処理によって定義した．被験者 i のデータで学習した分類器を被験者 j に適用したときの F1 を $s_{i\to j}$，逆方向である被験者 j で学習し被験者 i に適用したときの F1 を $s_{j\to i}$ とし，調和平均で平均識別率 S_{ij} を定義した．

$$S_{ij} = \frac{2s_{i\to j}s_{j\to i}}{s_{i\to j} + s_{j\to i}} \quad (1)$$

一方向のみ高精度な場合，学習データの分布の偏りや外れ値による影響が疑われるが，調和平均を用いることで，一方向だけ精度が高い組を過大評価せず，方向依存性を抑制する指標とした．この平均識別率を被験者間の類似度と見なし，平均識別率 S_{ij} から距離 D_{ij} を定義した．

$$D_{ij} = 1 - S_{ij} \quad (2)$$

全被験者ペアについて，D_{ij} を算出し，距離行列に対して階層クラスタリングを適用し，被験者の群構造を抽出した．平均識別率の上位 10 組を表 2 に示す．平均識別率の上位 10 組のうち，R を含む組は 4 組あり，R と近傍の高類似被験者が上位に含まれることが確認された．また階層クラスタリングのデンドログラムを図 4 に示す．またデンドログラムから読み取れる R を含む{A,F,I,R,S,U}の 6 名（全員男性，20 代 4 名，40 代 2 名）を一群として，精度検証を行った．

表 2　平均識別率の上位 10 組

rank	Subject1	Subject2	F1
1	A	R	0.712
2	I	R	0.697
3	L	R	0.695
4	I	S	0.689
5	H	R	0.675
6	B	L	0.672
7	I	U	0.671
8	N	P	0.669
9	F	S	0.662
10	B	D	0.659

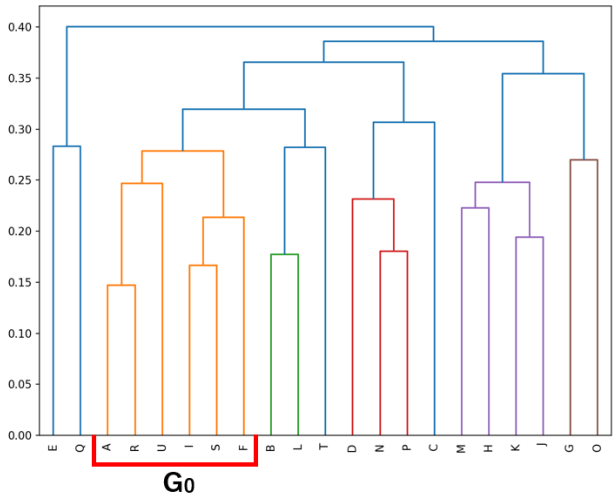

図 4　階層クラスタリングのデンドログラム

4.2. 精度検証

4.2.1. 個人差に着目した精度比較

個人差を考慮しない 21 名のデータをすべて含む全体グループG_{all}と 4.1 節で抽出した類似グループG_oの精度を比較することでクラスタリングによる個人差の影響を分析した．使用モデルとして SVM（kernel：rbf）を用いた．特徴量は 20 次元（AU17 次元，頭部運動 3 次元）である．被験者ごとの leave-one-out cross-validation(LOOCV)を実施し,クラスごとの Precision, Recall, F1 の値で精度比較を行った．被験者ごとの LOOCV では fold 間でテストサンプル数が異なるため，クラス別指標は各 fold における同意ラベルの件数，不同意ラベルの件数を重みとする加重平均で算出した．全体グループG_{all}と類似グループG_oの推定精度結果を表 3 に示す．結果として，両者を比較すると個人差を考慮する条件の方が同意ラベル，不同意ラベルの推定精度（F1）がそれぞれ 0.064，0.155 高くなった．特に不同意ラベルの推定精度に大きな差が見られた．

表 3　グループごとの精度比較

グループ	推定	Precision	Recall	F1
G_{all}	同意	0.499	0.612	0.544
	不同意	0.498	0.385	0.420
G_o	同意	0.602	0.625	0.608
	不同意	0.595	0.567	0.575

※各指標は加重平均をとった値

4.2.2. 使用特徴量の違いによる精度比較

先行研究との違いとして，特徴量に新たに頭部運動を追加したことが挙げられる．そのため頭部運動データの追加がどのように推定精度に影響を及ぼすかを分析した．特徴量パターンを AU＋頭部運動，AU 単体，頭部運動単体とし，G_oのグループにおける精度比較を行った．分析に使用したモデルは SVM（kernel：rbf）であり，被験者ごとの leave-one-out cross-validation（LOOCV）を行った．特徴量の違いによる推定精度結果を表 4 に示す．結果より，AU＋頭部運動と AU 単体が同程度の推定精度となった．一方で頭部運動単体は同意・不同意共に 0.500 を下回り明確に低い推定精度となった．

表 4　特徴量の違いによる精度比較

特徴量	推定	Precision	Recall	F1
AU+ 頭部運動	同意	0.602	0.625	0.608
	不同意	0.595	0.567	0.575
AU	同意	0.610	0.616	0.607
	不同意	0.588	0.574	0.575
頭部運動	同意	0.427	0.429	0.421
	不同意	0.441	0.451	0.440

※各指標は加重平均をとった値

5. 考察

まずクラスタリングの導入した結果，類似グループの不同意ラベルの推定精度が顕著に高い値を示した．これはクラスタ内で分布のずれが縮小し，それに伴う決定境界の取りこぼしが抑えられたためだと考えられる．またG_{all}では同意の Recall は 0.612，不同意の Recall は 0.385 と同意に予測が偏る傾向が確認されたが，G_oでは同意の Recall を維持しつつ，不同意の Recall が顕著に上昇した．このことから平均識別率に基づくクラスタリングが不同意の過小検出を減らす有効な手段であることが確認された．また平均識別率をもとに作成した類似グループの属性は全員男性であり，20 代が 4 名，40 代が 2 名と共通して男性であるという属性の類似点が確認された．したがって抽出された類似性には属性要因が反映されている可能性がある．

特徴量比較では AU＋頭部運動と AU 単体が同等の推定精度であり，頭部運動を追加したことによる精度向上は見られなかった．また頭部運動単体でも高精度の推定結果は得られず，単独特徴としても有効とは言えない．したがって，本研究の条件下では主要な特徴量は AU であり，頭部運動は単独でも AU と併用でも有意な結果を得ることができないことが確認された．本稿の分析において頭部運動に有効性が見られなかった原因として，今回用いた SVM は各フレームを単独サンプルとして扱うため，うなずきや横振り，首をかしげる動作などの頭部運動の時間構造を扱わないことが挙げられる．したがって時間微分や移動平均といった時間的特徴量を追加することで有効性が示される可能性がある．

6. 終わりに

本研究では同意・不同意のラベルが付与された 21 名のオンライン会議映像データセットを対象に平均識別率に基づくクラスタリングの導入，使用特徴量による推定精度の変化を分析した．

クラスタリングの導入により同意の推定精度を維持しつつ，不同意の推定精度が高くなった．また AU＋頭部運動と AU 単体使用時の推定精度には差が見られず，頭部運動単体を使用したときは同意・不同意の推定精度がいずれも 0.500 を下回った．これは時間的構造を考慮していなかったことが原因として挙げられる．

今後の展望として，時系列特徴量を用いて，AU および頭部運動の時系列情報の有効性を検証する．また本研究のデータセットは 21 名分に限られるが，クラスタリングを前提とする分析手順では更なるデータ数の確保が求められる．

7. 謝辞

本研究の一部は JSPS 科研費 JP21K12075 の助成を受けたものです．

文　献

[1] 総務省，「令和 6 年通信利用動向調査」，令和 7 年 5 月 30 日公表. [Online]. Available: https://www.soumu.go.jp/johotsusintokei/statistics/data/250530_1.pdf

[2] 高木幸子，“コミュニケーションにおける表情および身体動作の役割，” pp.25-35.

[3] 松井恵里香，“若年層における非言語行動の研究：若者の視線行動を中心として，” 昭和女子大学大学院日本語教育研究紀要, vol.2, pp.103‐109, 2004.

[4] C. O. Kruzic, D. Kruzic, F. Herrera, and J. Bailenson, “Facial expressions contribute more than body movements to conversational outcomes in avatar-mediated virtual environments,” *Scientific Reports*, vol.10, no.20626, pp.1–16, Nov.2020.

[5] Kang, H., Yang, R., Song, R., Yang, C., and Wang, W., An approach of query audience’s attention in virtual speech, Sensors, vol.24, no.16, 5363, 2024.

[6] A. Levordashka, D. Stanton Fraser, and I. D. Gilchrist, “Measuring real-time cognitive engagement in remote audiences,” Scientific Reports, vol.13, no.1, 10516, Jun.2023.

[7] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, “OpenFace 2.0: Facial Behavior Analysis Toolkit,” Proc. 13th IEEE Int. Conf. on Automatic Face & Gesture Recognition (FG 2018), pp.59–66, Xi’an, China, 2018.

[8] P. Ekman and W. V. Friesen, Facial Action Coding System: A Technique for the Measurement of Facial Movement, Consulting Psychologists Press, Palo Alto, CA, 1978.

[9] 三浦辰志，時枝大貴，岡田将吾，“メタ学習手法を用いた対話時のユーザーの感情推定，”2025 年度人工知能学会全国大会論文集,May 2025.

[10] 齊藤寛己，佐藤矯汰郎，大和淳司，“オンライン会議における顔画像による同意，不同意の推定に有効な特徴量の検討，”信学技報（HCS），vol.123, no.241–242, pp.45–49, Nov.2023.

[11] H. Sloetjes and P. Wittenburg, “Annotation by category – ELAN and ISO DCR,” Proc. Sixth Int. Conf. on Language Resources and Evaluation (LREC 2008), pp.816–820, Marrakech, Morocco, May 2008.