

話者の頭部及び姿勢変化に着目した複数人対話分析

小松 和朗^{1,a)} 嶋田 和孝^{2,b)} 遠藤 勉^{2,c)}

概要: 多人数インタラクションの円滑化において話者の内面を推定することは重要な課題である。本論文では話者の頭部及び姿勢変化のタイミングと話者の内面変化との関連性に着目した複数人対話分析を行う。提案手法では、頭上方向から撮影した人物画像より得られた深度情報を基に抽出した (1) 深度ヒストグラム, (2) 人物の面積, (3) 身体の傾き度合い, (4) 身体の向き の計 4 種類の特徴量を用いる。画像フレーム間において、取得した各特徴量の類似度を比較することで話者の頭部及び姿勢変化推定を行う。また、対話分析では話者の姿勢変化と感情変化の発生回数に着目し、調査することで得られた得た見解について述べる。

Multi-party interaction analysis based on the speaker's changes in head and posture

KAZUAKI KOMATSU^{1,a)} KAZUTAKA SHIMADA^{2,b)} TSUTOMU ENDO^{2,c)}

Abstract: The goal of our study is to understand user's states in multi-party conversation. In this paper, we focus on the change of a user's posture in the conversation. Our method uses a top view image with depth information for the analysis. We compute a similarity between each image frame by using 4 features; (1) depth histogram, (2) dimensions, (3) slant and (4) direction. Furthermore, we analyze our corpus in terms of the relation between the changes of posture and emotion of users.

1. はじめに

対話は、私たち人間が生活する上で、他人とコミュニケーションを図るための手段である。コミュニケーションは、自らの意思を相手に伝え、相手の意思を受け入れることで成立し、他人との情報共有、双方の感情理解および全体の意思決定を目的として行われる。雑談や会議といった複数人で行うコミュニケーションの形態は多人数インタラクションと呼ばれ、現在対話分析を中心に研究が進められている [1]。我々は複数人対話において言語情報および非言語情報を効果的に組み合わせて使用することで話者の内面

を推定し、推定結果を基に各話者の意図伝達、意図理解を促進するインタラクション支援システムの構築を目指している。これまでに、単語の出現頻度や発話長、声の抑揚や発話時間を用いた複数人対話における盛り上がり判定 [2] や、対話中に発生する笑いに着目し、笑いの発生要因や笑いの大きさに基づく盛り上がり推定手法 [3] について研究を進めてきた。これらの研究においてより精度を向上させるには、画像情報の利用が必要である。対話中の画像情報から取得できる特徴としては話者の表情、顔の向きや姿勢、ジェスチャなどが挙げられるが、我々は対話中に話者がみせる姿勢に着目する。

本研究では話者が対話中にみせる姿勢とその時の話者の内面変化との関連性に着目した対話分析を行う。頭部の動きや姿勢は話者の内面状態を無意識のうちに表現する特徴の 1 つである。例えば、相手を注視しつつタイミングよく相槌をうてば、話題に関心を持っている状態であるといえるし、相手を見ておらず頬杖をついていたならば話題に関心がない状態であるといえる。我々は「人の姿勢パターン

¹ 九州工業大学大学院情報工学府
Kyushu Institute of Technology, 680-4 Kawazu, Iizuka-shi,
Fukuoka, 820-8502, JAPAN

² 九州工業大学情報工学部
Kyushu Institute of Technology, 680-4 Kawazu, Iizuka-shi,
Fukuoka, 820-8502, JAPAN

a) k_komatsu@pluto.ai.kyutech.ac.jp

b) shimada@pluto.ai.kyutech.ac.jp

c) endo@pluto.ai.kyutech.ac.jp

はある程度一定であり、姿勢パターンが変化した時に話者の内面に変化が発生する」という仮説を検証することで、話者の頭部及び姿勢変化が話者の内面推定及び対話分析に対して有効であることを示す。

2. 関連研究

本節では、対話分析における非言語情報の有効性及び姿勢変化の重要性、頭上方向から撮影するメリットについて述べた関連研究をそれぞれ紹介する。Vargas[4] は対話中に現れる姿勢や視線が話者の内面推定に有効な要素である言語調整動作であることを述べている。Mahmoud ら [5] は自発的に生じたジェスチャとその時の精神状態との関係性を考慮した感情推定手法を提案しており、人物の感情認識における非言語情報の有効性を示している。

非言語情報のうち、頭部の動きや姿勢は対話分析や学習支援の分野において学習者の内面推定を行う際にその有効性が述べられている。頭部の動きに関する研究として、熊野ら [6] は複数人対話において各話者の表情と話者間の視線から共感状態を求めることで各話者の対人感情推定を行っており、円滑な多人数インタラクションの実現に向けた新しい研究の枠組みを提案している。また、姿勢に関する研究として、Mota ら [7] は「学習者の内面が姿勢に現れる」という仮定に基づき、姿勢推定を行うことで学習者の内面推定を図っている。Grafsgaard ら [8] は深度センサーを学習者の正面に設置し、学習者の各部位の情報を間接的に取得することで姿勢推定を行い、姿勢と心理状態との相関関係を算出している。

しかし、これらの研究にはいくつかの問題点が存在する。Mota らの研究では姿勢推定に用いる特徴量を圧力センサーを搭載した椅子から物理的に取得した圧力値から算出しており、椅子と接触した部位の情報を詳細に取得できる反面、腕や手の動きなど接触していない部位の情報を取得することはできない。加えて、複数人対応を実現するためには人数分の装置を用意する必要があるなど制約が多い。Grafsgaard らの研究では人物や身体の一部の重なりによって身体の一部にオクルージョンが生じた場合、取得できる情報量が減少するため識別精度に悪影響を及ぼすことが予想される。対話分析手法においても中村ら [9] は、対話のトピックや話者間の関係性を考慮する必要があると述べており、単純に視線や姿勢状態と感情を一対一対応させればよいという問題ではない。

以上の関連研究を踏まえて、本研究では深度センサーを備えたカメラを用いて頭上方向から撮影することで話者の姿勢状態及び複数人対話環境を取得する。取得した動画データより、話者の姿勢状態及び姿勢変化のタイミングを推定するとともに複数人対話環境における対話分析を行う。また、我々は人物識別のタスクでオクルージョンを解消するために頭上画像を利用する手法を提案している [10]。頭

上方向から撮影することで人物の重なりによって発生するオクルージョンを回避することができ、かつ、被験者に与える心理的負担の軽減が期待できる。さらに、撮影に深度センサーを利用することで物体との距離情報を含んだ深度情報を取得することができ、撮影時に照明変動の影響を受けないというメリットがある。

3. コーパス

3.1 データ収集

本論文では3人1グループで行う複数人対話を対象としている。そこで、今回は計19名の大学生を対象に対話者として実験に参加してもらい、複数人対話データを8グループ分、計16セット収集した*1。

実際に撮影した動画データの一部分を図1に示す。1つのテーマにつき4分間の対話をしてもらい、各グループ2テーマ分の対話データを収集している。対話のテーマは「研究室」、「就職活動」といった対話者全員が同程度の知識を持ちうる事柄にしている。対話データを録画する直前に対話のテーマを提示することでなるべくテーマに沿った対話を行ってもらうようにした。また、各話者には実験の目的を説明していない。これは、実験において話者に特別な意識をもたせないためであり、実際の対話環境における自然な動作や表情を取得するためである。なお、本実験では動画撮影にASUS社が提供しているXtion PRO LIVE*2を使用している。

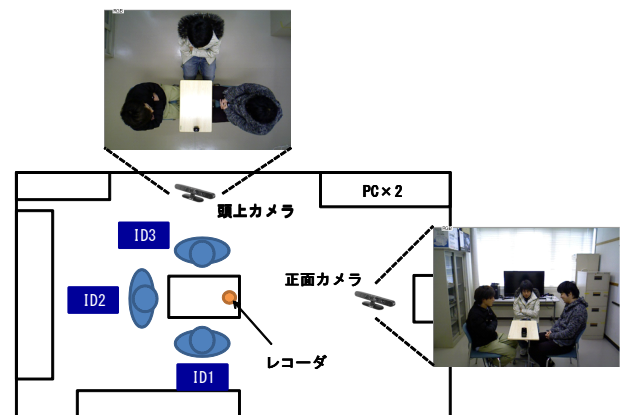


図1 姿勢変化推定及び対話分析における撮影環境。

3.2 フレームの正規化

今回の実験では正面及び頭部方向に設置した深度センサーをそれぞれ同程度のPC環境で利用することで動画データを取得したが、正面及び頭上方向で取得された動画

*1 机上に設置したレコーダにより動画データと同時に音声データも取得したが、本研究では頭部及び姿勢変化に着目した対話分析を行うため今回は使用していない。

*2 ASUS Xtion PRO LIVE, http://www.asus.co.jp/Multimedia/Motion_Sensor/Xtion_PRO_LIVE/

データのフレーム数に若干の差分が生じてしまった。このままではそれぞれの深度センサーで取得したフレーム番号同士の対応付けがとれていないため単純に比較することができない。そこで、正面方向のフレームを F 、正面方向の総フレーム数を F_{center} 、頭上方向の総フレーム数を F_{top} として、正規化後の正面方向のフレーム F_{normal} を以下の式によって正規化を行った。

$$F_{normal} = F \times \frac{F_{top}}{F_{center}} \quad (1)$$

この正規化によって算出された新たな正面方向のフレーム F_{normal} は頭上方向のフレーム数と対応付けすることができる。

3.3 姿勢変化のアノテートデータ

収集した対話データに対し、頭部及び姿勢変化、感情変化が発生したフレームのラベル付けを行った。ラベル付けは、3名のアノテータによって行った。3名のうち、1名は実際に対話実験に参加してもらっており、残り2名は実験には参加していない。各アノテータには各対話者の頭部方向・姿勢・感情の変化についてそれぞれ評価してもらった。頭部方向及び姿勢変化は変化開始時と変化終了時のフレームをそれぞれ記録してもらったが、感情変化は変化開始時のフレームのみを記録してもらった。これは感情の変化が頭部の動きや姿勢とは異なり、一瞬のうちに発生するためである。また、ここでいう「姿勢変化が発生した」という状態とは対象者の身体の一部が明らかに動いていると人目で判断できる状態としており、首ふりやうなずきといった微かな動作に関しては今回対象外としている。

4. 頭部及び姿勢変化推定

4.1 提案手法概要

本節では本研究における提案手法の概要について述べる。提案手法は深度情報を利用した人物領域の抽出、特徴量抽出、頭部及び姿勢変化推定の大きく3つによって構成される。

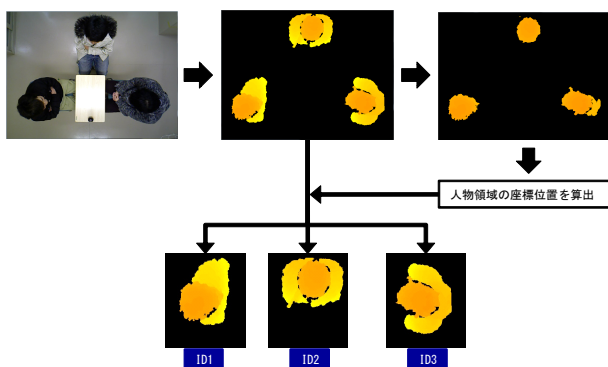


図2 人物領域の抽出。

人物領域の抽出

姿勢状態推定では人物領域の抽出対象が一人であるのに対し、姿勢変化推定では図2で示すように抽出対象が複数人物であるため、話者毎の人物領域を抽出する必要がある。抽出には各話者の頭部領域を算出し、頭部領域の座標を基に人物領域の抽出範囲を決定している。なお、各話者の人物領域の範囲は 213×240 としている。

特徴量抽出

本論文では、話者の姿勢変化を取得するために人物領域より取得した深度情報を基に以下の4種類の特徴量を抽出している。

1. 深度ヒストグラム特徴
2. 人物の面積特徴
3. 身体の傾き度合い特徴
4. 身体の向き特徴

頭部及び姿勢変化の推定

頭部及び姿勢変化の推定はまず、姿勢変化が発生したフレームから前述した4つの特徴量を取得する。次にフレーム毎において各特徴量を抽出し、前回姿勢変化が発生したフレームで抽出された特徴量との類似度を算出する。この時得られた類似度を基に閾値を設定し、各フレームにおいて閾値以上の差分が見られた時にそのフレームを姿勢変化フレームとして判断する。図3に変化推定の例を示す。

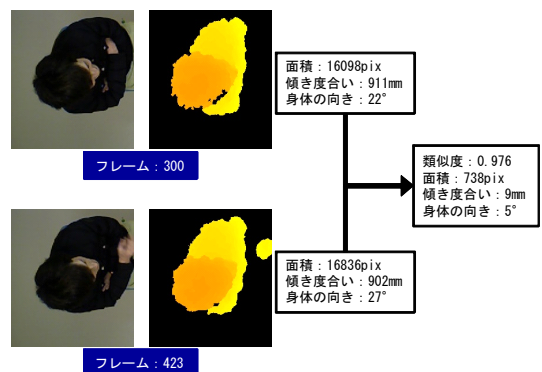


図3 頭部及び姿勢変化推定の例。

4.2 深度ヒストグラム特徴

4.2.1 深度ヒストグラムによる特徴抽出

姿勢状態推定の手法では、人物領域のみを取得するために最適な距離情報の範囲を $1000\text{mm} \sim 1700\text{mm}$ と設定することで深度ヒストグラムを作成している。しかし、 $1000 \sim 1300\text{mm}$ の範囲では他の範囲と比較して話者の人物領域が現れる回数が少なく、ヒストグラムに特徴が現れにくい。そこで、範囲ごとにヒストグラムを構成する要素の基準を差別化したり、ある一定の範囲において要素を統合するこ

とで特徴の顕著化を図る必要がある．本研究では 1000～1300mm の範囲において 10mm ごとに、1301～1700mm の範囲においては 1mm ごとにヒストグラムの要素を生成しており、合計 430 個の要素でヒストグラムを構成している．図 4 に深度ヒストグラムの生成例を示す．

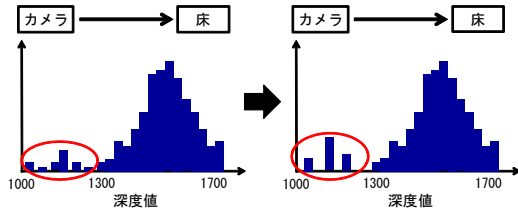


図 4 ヒストグラム生成の例．

4.2.2 深度ヒストグラムによる類似度算出

深度ヒストグラム特徴の類似度算出には深度ヒストグラム間の Bhattacharyya 距離を用いる．ヒストグラム間の Bhattacharyya 距離による類似度を算出するには、それぞれのヒストグラムでヒストグラムの合計が 1 になるように正規化を行う． m ビンからなる 2 つの正規化深度ヒストグラムをそれぞれ p, q とし、前回変化が発生したフレームの深度ヒストグラムを p 、判定対象フレームの深度ヒストグラムを q 、各ビンを p_u, q_u とする．このときの 2 つの正規化深度ヒストグラムの類似度 S_{hist} は以下の式で算出される．

$$S_{hist} = \sum_{u=1}^m \sqrt{p_u q_u} \quad (2)$$

4.3 人物の面積特徴

4.3.1 人物の面積による特徴抽出

人物の面積は姿勢変化推定においても直感的で特有の特徴量となると考える．そこで本手法でも姿勢状態推定の手法と同様に抽出された人物領域のピクセル数の総和を算出することにより求める．

4.3.2 人物の面積による類似度抽出

人物の面積特徴の類似度算出には前回変化が発生したフレームの人物の面積特徴との差分を用いる．前回変化が発生したフレームの面積特徴を S_{n-1} 、判定対象フレームの面積特徴を S_n とする．このときの 2 つの面積特徴の類似度 S_{area} は以下の式で算出される．

$$S_{area} = |S_n - S_{n-1}| \quad (3)$$

4.4 身体の傾き度合い特徴

4.4.1 身体の傾き度合いによる特徴抽出

取得した画像データを分析した結果、姿勢状態推定の手法と同様に各姿勢の身体の傾き方には一定の傾向が現れていることが分かった．そこで、本研究においても人物領域の中心座標と頭部領域の中心座標を利用した身体の傾き度

合いによる特徴抽出を行う．

本論文では、画像モーメントを利用することで人物領域より取得した重心及び深度値を中心座標として扱う．人物の面積および体型特徴を取得した際に用いた人物領域を入力画像とし、以下の式において人物領域の 3 次の空間モーメント M_{pq} を求めることで領域の重心座標 $(\bar{x}, \bar{y})$ を算出する．式 (4) の p, q はそれぞれ x 方向、 y 方向の次数を示している．

$$\begin{cases} m_{pq} = \sum_{x,y} (x^p \cdot y^q \cdot f(x,y)) \\ \bar{x} = \frac{m_{10}}{m_{00}}, \quad \bar{y} = \frac{m_{01}}{m_{00}} \end{cases} \quad (4)$$

さらに、式 (4) で求めた座標地点の深度値を要素として加えることで 3 次元の座標として取得する．また、頭部領域の中心座標も同様に頭部領域を入力画像として式 (4) を用いることで頭部領域の重心座標を算出する．頭部の位置を示す頭部領域は図 5 において人物領域内の濃淡が濃い部分である．なお、図 5 中の ● は人物領域の中心座標点を表しており、▲ は頭部領域の中心座標点を表している^{*3}．入力画像より人物領域の中心座標からみて頭部領域の中心座標がどこに位置するか求めた後、座標間の距離を取得する．この座標間の距離の値を身体の傾き度合いとする．図 5 の矢印がその傾きを意味している．身体の傾き度合いを表す座標間の距離 D は以下の式で算出する．

$$D = \sqrt{(x_c - x_g)^2 + (y_c - y_g)^2 + (Depth_c - Depth_g)^2} \quad (5)$$

式 (5) の x_c と y_c はそれぞれ人物領域の中心座標 (x_c, y_c) を、 x_g と y_g はそれぞれ頭部領域の中心座標 (x_g, y_g) を、 $Depth_c$ と $Depth_g$ はそれぞれ人物領域の中心座標と頭部領域の中心座標の深度値 $Depth_c, Depth_g$ を表現している．図 5 に身体の傾き度合い特徴の取得例を示す．図中の矢印の長さが身体の傾き度合いである座標間の距離 D を表している． D の値が大きければ大きいほど矢印の長くなり、身体が大きく傾いていることを表す．

4.4.2 身体の傾き度合いによる類似度抽出

身体の傾き度合いの類似度算出には前回変化が発生したフレームの身体の傾き度合い特徴との差分を用いる．前回変化が発生したフレームの傾き度合い特徴を D_{n-1} 、判定対象フレームの傾き度合い特徴を D_n とする．このときの 2 つの面積特徴の類似度 S_{dist} は以下の式で算出される．

$$S_{dist} = |D_n - D_{n-1}| \quad (6)$$

4.5 身体の向き

4.5.1 身体の向きによる特徴抽出

姿勢状態推定の手法では人物領域の中心座標 (x_c, y_c) お

^{*3} 図 5 の画像で取得する座標位置は実世界の寸法 (単位は mm) に変換している．図 5 の身体の傾き具合を取得する際に使用する画像のみが対象であることに注意．

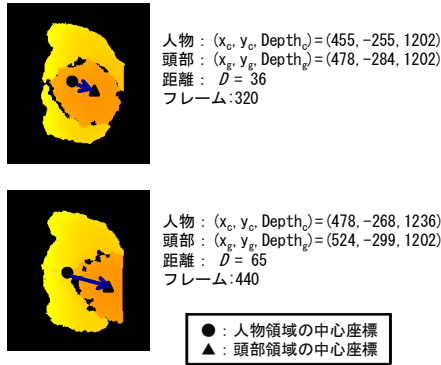


図 5 人物領域および頭部領域の中心座標と
身体の傾き度合い特徴の例。

よび頭部領域の中心座標 (x_g, y_g) を用いて角度を算出することで身体の向きは算出していた。本論文では、身体の向きの算出に頭部領域における楕円検出手法を利用している。楕円検出には最小二乗推定を用いており、最良の楕円及び楕円の中心座標を算出することができる。さらに、図 6 において楕円の中心座標 (x_1, y_1) および楕円の短辺 a における等分点座標 (x_2, y_2) を用いて身体の向きとなる角度を算出する。身体の向き θ は以下の式で算出する。

$$\theta = -\arctan((y_2 - y_1), (x_2 - x_1)) * (180/\pi) \quad (7)$$

なお、座標位置は図 6 の画像において、その左上を $(0,0)$ とした絶対座標である。図 6 は身体の向きの取得例を表している。

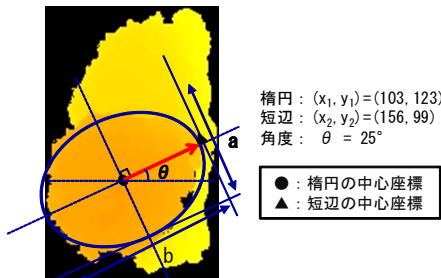


図 6 身体の向き特徴の例。

4.5.2 身体の向きによる類似度抽出

身体の向きの類似度算出には前回変化が発生したフレームの身体の向き特徴との差分を用いる。前回変化が発生したフレームの身体の向き特徴を θ_{n-1} 、判定対象フレームの身体の向き特徴を θ_n とする。このときの 2 つの面積特徴の類似度 S_{angle} は以下の式で算出される。

$$S_{angle} = |\theta_n - \theta_{n-1}| \quad (8)$$

4.6 変化推定における閾値設定

4.2.2 節より得られた深度ヒストグラム特徴による類似度 S_{hist} 、4.3.2 節より得られた人物の面積特徴による類似度 S_{area} 、4.4.2 節より得られた身体の傾き度合い特徴による類似度 S_{dist} 、4.5.2 節より得られた身体の向き特徴による

類似度 S_{angle} の 4 つの類似度に対してそれぞれ閾値を設定した。事前に行った予備実験の結果、各類似度が $S_{area} \geq 200\text{pix}$ かつ $S_{dist} \geq 5\text{mm}$ を満たした上で、 $S_{hist} \leq 0.985$ または $S_{angle} \geq 10^\circ$ になった状態が変化推定に最適であると判明したため、本論文においても同様に設定している。

4.7 評価実験

4.7.1 評価基準

提案手法で取得したデータとアノテートデータとの比較を行うにあたり、本論文では ± 1 秒間を誤差の範囲内とし、アノテートデータに記載された各フレームの前後 20 フレームを正解フレームとしている。これは 1 フレームが約 0.05 秒という人間が変化を捉えるには非常に短い時間であり、同一フレームを取得するのは非常に困難なためである。また、本研究では変化推定の精度より対話分析に重点を置いている。そのため、1 秒という時間は人間が刹那と感じる時間の範囲内であり、その区間の変化は人間にとっては気にならない誤差であると考えた。なお、この条件で 3 名分のアノテートデータの一致率を算出した結果、頭部及び姿勢変化における一致率は 86.5 % となった。以上を踏まえ、本論文では以下の 3 項目を評価することで提案手法の有効性を検証した。

1. データ要約率
2. 抽出成功率
3. 正答率

評価内容の概略を図 7、各評価項目間の関係性を図 8 に示す。まず、提案手法によって対話データから姿勢変化フレームと推定したフレームを取得する。次に、推定したフレームとアノテータデータを直接比較するとともに、推定したフレームを人間の評価尺度に変換する。これは、提案手法のシステムとしての評価尺度がフレーム単位である一方、人間の評価尺度は変化回数単位であり、提案手法を対話分析に応用する際に人間とシステム間にギャップが生じてしまうためである。最後に、人間の評価尺度に変化したものとアノテートデータを比較する。システムの評価尺度から人間の評価尺度に変換することにより、システムと人間両方の評価尺度で有効性を評価することができる。

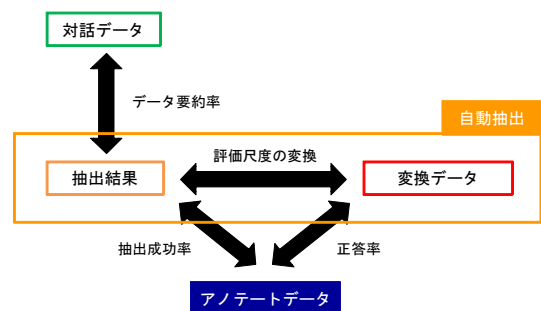


図 7 評価内容の概略図。

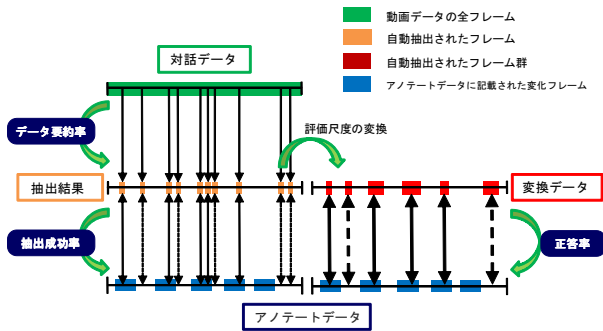


図 8 各評価項目間の関係性。

データ要約率

一般的に対話分析を行う際には、収集した動画データを繰り返し観察する必要があるため非常に大きな労力を要する。そこで、提案手法を用いて分析に必要な姿勢変化が発生したフレームを残しつつ分析に不要なフレームが削減できれば、今後対話分析を効率良く行う上で大きな貢献ができる。データ要約率とは各グループの対話データの全フレームから提案手法によって抽出されたフレーム数の割合であり、以下の式によって算出する。

$$\text{データ要約率 (\%)} = \frac{\text{抽出フレーム数}}{\text{全フレーム数}} \quad (9)$$

データ要約率の値が高いほど対話分析で確認するフレームが少なくなるため、効率良く分析を行うことができる。

抽出成功率

次に抽出したフレームがどれだけ有効であるかを調べることで提案手法の有効性を検証した。ここでの正解フレームとは、抽出されたフレームのうちアノテートデータに記載された各フレームの前後 20 フレームのことを指す。提案手法によって取得された抽出フレームに対する正解フレーム数の割合を抽出成功率として提案手法の性能評価を行う。抽出成功率は以下の式によって算出する。

$$\text{抽出成功率 (\%)} = \frac{\text{正解フレーム数}}{\text{抽出フレーム数}} \quad (10)$$

図 8 において、二重線の矢印は正解フレームを、破線の矢印は失敗フレームを表しており、この場合、自動抽出された 10 フレームの内、アノテートデータに記載されたフレーム群と一致するフレームは 6 フレームであるため、抽出成功率は 60.0 % となる。抽出成功率は提案手法をシステムの評価尺度で評価することで取得した変化推定の精度を表しており、値が高いほど分析に必要な姿勢変化フレームが取得できていることを示す。

正答率

最後に、提案手法で人間の評価尺度において有効であるフレームがどの程度抽出されたかを調べることで提案手法の有効性を検証した。提案手法では全てのフレームに対して設定した制約に従って機械的に評価する一方で、人間は

連続して発生する姿勢変化を一つの姿勢変化として評価を行う。例えば図 8 では、自動抽出されたフレームのうち 6 フレームが姿勢変化推定に成功しているが、人間の評価尺度では 4 箇所の姿勢変化推定を成功したことになり、評価結果が異なってしまう。そこで、提案手法によって抽出された抽出フレーム数を人間と同じ評価尺度に変換したものを集約されたフレーム群と定義する。

さらに、集約されたフレーム群の内アノテートデータに記載されたフレーム群と一致するものを正解フレーム群、アノテートデータに記載されたフレーム群の合計を正解変化数と定義して、正解変化数に対する人間の評価尺度で評価された正解フレーム群の割合を正答率として評価する。正答率は以下の式によって算出する。

$$\text{正答率 (\%)} = \frac{\text{正解フレーム群}}{\text{正解変化数}} \quad (11)$$

図 8 の集約されたフレーム群において、二重線の矢印は正解フレーム群を、破線の矢印は失敗フレーム群を表しており、この場合、正解フレーム群は 4 つ、アノテートデータに記載されたフレーム群は 5 つあるため、正答率抽出成功率は 90.0 % となる。

正答率は提案手法を人間の評価尺度で評価することで取得した変化推定の精度を表しており、値が高いほど提案手法がどれだけ人間の感覚に近似した状態で分析に必要な姿勢変化のタイミングを捉えられているかを示す。

4.8 結果と考察

データ要約率、抽出成功率、正答率に関する実験結果をそれぞれ表 1 に示す。まず、8 つの対話データにおけるデータ要約率の平均は 91.7 % となった。次に、抽出したフレーム数における正解フレームの抽出成功率の平均は 76.4 % と、分析に不要なフレームを含んでいるものの一定の成果を挙げていることが分かる。最後に、アノテータデータに記載されたフレーム数における人間の評価尺度に変換した提案手法の正答率の平均が 94.4 % であることから、提案手法によって対話分析に必要な姿勢変化フレームを人間の感覚に近似した状態で効率良く取得できたことが分かる。一方、提案手法では図 9 で示すような机と同等の位置または机の下で発生した姿勢変化を捉えることができなかった。捉えられなかった原因としては、本研究で設定した深度センサーの範囲内に机周辺の情報が含まれておらず、その結果深度情報を取得することができなかったことが考えられる。

5. 対話データ分析

本章では、3.1 節で収集した対話データにおけるアノテートデータの分析を行う。本研究では以下の 2 項目を分析及び評価を行った。

表 1 データ要約率.

	全フレーム数	抽出フレーム数	正解フレーム数	正解変化数	正解フレーム群	データ要約率	抽出成功率	正答率
Group1	27900	1827	1512	390	369	93.5 %	82.8 %	94.6 %
Group2	27735	2613	2255	506	482	90.6 %	86.3 %	95.3 %
Group3	28545	1737	817	231	231	93.9 %	47.0 %	97.9 %
Group4	29154	3238	2850	537	537	88.9 %	88.0 %	92.6 %
Group5	30600	3536	1991	412	412	88.4 %	56.3 %	100.0 %
Group6	21945	1974	1869	342	342	91.0 %	94.7 %	96.6 %
Group7	27687	1249	1109	307	307	95.5 %	88.8 %	83.0 %
Group8	32731	2532	1890	457	457	92.3 %	74.6 %	96.4 %
合計	226297	18706	14293	3322	3137	91.7 %	76.4 %	94.4 %

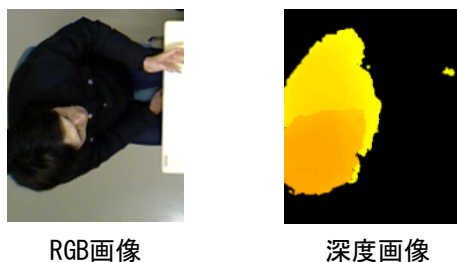


図 9 姿勢変化フレーム取得失敗例.

1. 姿勢変化と感情変化の回数
2. 姿勢変化と感情変化の一致率

5.1 姿勢変化と感情変化の回数

まず初めに、3 節で述べたアノテートデータにおいて頭部及び姿勢変化と感情変化の発生回数に着目した分析を行った^{*4}. アノテートデータの一部を表 2 に示す. その結果、姿勢変化と感情変化の回数には以下の 3 項目との関連性がみられた.

1. 話者の特性推定
2. 話者間の親密度
3. 対話場の雰囲気推定

表 2 アノテートデータ 1 の姿勢変化と感情変化の回数.

	姿勢変化数				感情変化数				全変化数
	ID1	ID2	ID3	合計	ID1	ID2	ID3	合計	
Group1	56	6	52	114	17	21	20	58	172
Group2	48	42	56	146	14	27	14	55	201
Group3	6	20	34	60	12	7	5	24	84
Group4	54	50	72	176	14	34	22	70	246
Group5	36	48	84	168	27	30	28	65	253
Group6	18	80	38	136	10	24	20	54	190
Group7	58	36	42	136	14	17	13	44	180
Group8	72	62	44	178	27	34	29	90	268

5.1.1 話者の特性推定

表 2 において、Group4 の ID3 や Group5 の ID3 のような他の話者に比べて頭部及び姿勢変化が多かった話者は、対話場における中心人物となっている傾向が確認できた.

^{*4} 感情変化における 3 名分のアノテートデータの一致率を姿勢変化と同様に 4.7.1 節で述べた評価基準で算出した結果、一致率は 79.3 % となった.

これは話し手として話題を盛り上げようとする際に、周囲の反応を伺いつつ、話題に惹きつけるための動作を加えて話をしているためである. 一方で、Group2 の ID2 や Group3 の ID3 のように頭部及び姿勢変化が少なく、かつ、感情変化が最も多かった話者は、話の聞き手にまわる傾向が確認できた. これは現在の話題が自分の前提知識に無い話題であり、話し手から情報を得るために聞き手に徹していたことが分かる. このことから話者の姿勢変化の回数は、話者の特性を判断するための有効であることが分かった.

5.1.2 話者間の親密度

表 2 において、Group4 や Group5、Group8 のような姿勢及び感情変化が多いグループは、話者の一人一人が相手の顔を見ながら話したり、ジェスチャを交えながら話す様子が多く見られ、会話がほとんど途切れることがなかった. 一方で、Group3 のように変化が少ないグループでは、相手の顔を見ずに下を見ながら話したり、長時間沈黙が続く傾向が見られた. これは話者同士の親密度が高いグループほど相手の好みや話題に対する知識量がどの程度であるかを把握しており、相手のペースに合わせつつ話題を展開出来る一方、親密度が低いグループでは相手の好みや話題に対する知識量が分からないため、相手の状態を探りながら話題を展開せざるを得ないことが影響していると考えられる. このことから、姿勢及び感情の変化回数がグループの親密度を図る指標の一つになりうるということが分かった.

5.1.3 対話場の雰囲気推定

表 2 において、Group5 のような頭部及び姿勢変化、感情変化が多く、かつ各話者の変化回数がほぼ同じグループでは、話者間でのやりとりが多く、場が非常に盛り上がっている様子が確認出来た. 一方で、Group3 のように頭部及び姿勢変化、感情変化が少ないグループでは、話者間のやりとりがほとんどなく、場があまり盛り上がっていない様子が確認出来た. 対話場が盛り上がっているグループでは各話者が話し手と聞き手の立場を無意識のうちに交代しながら対話を行っているため、話者全員の姿勢及び感情変化の回数が多いことが考えられる. 一方で、対話場が盛り上がっていないグループでは話者間で話し手と聞き手の役割を分担できないまま対話のやりとりを試みており、上手

く対話をできない状態が続いたため、話者全員の姿勢及び感情変化の回数が少なくなっていることが考えられる。このことから、姿勢及び感情の変化回数が対話場の雰囲気や推定する上で有効であるということが分かった。

姿勢変化のパターン

前述の3項目に加えて対話において特定の話者に共通した動作がいくつかが見られた。まず、メガネを着用している話者に共通してみられた動作として話者が話終わった直後にメガネの位置を直すしぐさである、また、対話場における中心人物である話者は話題に詰まった時や話をする直前に上を見上げるしぐさを多くすることが確認できた。さらに、場が盛り上がっていない時には話者全員が下を向いていた。このように対話においていくつかの行動パターンが存在していると考えられ、今後の対話分析において着目すべき要素である。

5.2 姿勢変化と感情変化の一致率

次に、話者の頭部及び姿勢変化と感情変化との関連度を調査するために、姿勢変化と感情変化の発生箇所の一致率を測定した。アノテートデータの一部を表3に示す。その結果、一致率はアノテータ3名の平均で32.1%となり、話者の感情変化の際に必ずしも動作が発生するわけではないということが明らかとなった。

表3 アノテートデータ1の姿勢変化と感情変化の一致率。

	変化総数	姿勢変化数	感情変化数	一致数	一致率
Group1	172	114	58	14	24.1 %
Group2	201	146	55	22	40.7 %
Group3	84	60	24	3	12.5 %
Group4	246	176	70	44	62.9 %
Group5	253	168	65	40	47.1 %
Group6	190	136	54	28	51.9 %
Group7	180	136	44	21	47.7 %
Group8	268	178	90	48	53.3 %
合計	1594	1114	479	220	45.9 %

考えられる要因として以下の2点が考えられる。まず一つ目は、話者によって自己の感情表現が異なるという点である。実際の動画データを分析した結果、話が盛り上がっている際に身体を大きく揺らしながら笑う人もいれば、身体をほとんど動かさずに笑う人もいるなど、話者によって感情の表現が異なっていたことが分かった。そのため、話者全員の姿勢及び感情変化を一律で評価するのではなく、話者毎に一定の評価パターンを設ける必要がある。二つ目は、話者の何名かがメガネやマスクを着用していた点である。メガネ及びマスクの着用によって話者の表情にオクルージョンが生じたことで、人手においても感情変化の推定が困難な状態であった。そのため、顔の一部分である目や口元に重点をおいて推定を行うなど、顔の表情に加えてオクルージョンに頑健な特徴量を用いた感情推定を行う必要がある。

6. おわりに

本論文では話者が対話中にみせる姿勢とその時の話者の内面変化との関連性に着目した対話分析を行った。実験の結果、対話データ全体のフレーム数を平均91.7%削減しつつ、姿勢変化が発生したフレームを人間の評価尺度において平均94.4%の精度で抽出することができた。このことにより、提案手法が対話分析において有効であることが示された。さらに対話分析の結果、話者の姿勢及び感情変化が話者の特性判定や親密度、場の雰囲気推定を行う上で有効な特徴量であることが示された。今後の課題としては、最終目標である対話支援の実現にむけたシステム設計に加え、人物領域検出の改良や、変化発生を捉えるための特緒量の追加が必要である。

参考文献

- [1] 坊農真弓, 高梨克也, “多人数インタラクションの分析手法,” オーム社, 2009.
- [2] 横山貴彦, 嶋田和孝, 遠藤 勉, “複数人談話における言語情報と非言語情報を利用した盛り上がり判定,” 言語処理学会第18回年次大会 (NLP2012), 1-26, 2012.
- [3] 嶋田和孝, 楠本章裕, 横山貴彦, 遠藤 勉, “複数人談話における笑いの情報を考慮した盛り上がり判定,” 電子情報通信学会, 言語理解とコミュニケーション研究会 (NLC), NLC2012-7, pp.25-30, 2012.
- [4] Marjorie Fink Vargas, 石丸 正, “非言語コミュニケーション,” 新潮社, 1987.
- [5] M. Mahmoud, T. Baltrusaitis, P. Robinson and L. Riek, “3D corpus of spontaneous complex mental states,” In Proceedings of the International Conference on Affective Computing and Intelligent Interaction (ACII 2011), Lecture Notes in Computer Science, Oct, 2011.
- [6] 熊野史朗, 大塚和弘, 松田昌史, 大和淳司, “二者間の行動の一致性と時間差に基づく外部観察者の共感解釈の分析,” 電子情報通信学会, ヒューマンコミュニケーショングループ (HCG), HCG2012-I-7-3, pp.330-337, 2012.
- [7] S. Mota and R. W. Picard, “Automated Posture Analysis for detecting Learner’s Interest Level,” In Proceedings of Workshop on Computer Vision and Pattern Recognition for Human-Computer Interaction, CVPR HCI, June, 2003.
- [8] J. F. Grafsgaard, K. E. Boyer, E. N. Wiebe and J. C. Lester, “Analyzing Posture and Affect in Task-Oriented Tutoring,” In Proceedings of the 25th Florida Artificial Intelligence Research Society Conference, 2012.
- [9] 中村和晃, 角所 考, 正司哲朗, 美濃導彦, 澤木美奈子, 南泰浩, 前田英作, “擬人化エージェントとの音声対話時におけるユーザの非言語動作からの難/易及び興味/退屈の推定,” 電子情報通信学会和文論文誌 A, Vol.J95-A No.1 PP.85-96, 2012-1.
- [10] R. Nakatani, D. Kouno, K. Shimada and T. Endo, “A Person Identification Method Using a Top-view Head Image from an Overhead Camera,” In Proceedings of 2nd International Workshop on Advanced Computational Intelligence and Intelligent Informatics (IWACIII2011), SS4-1, pp.739-746, 2011.