J-007

画像処理技術を用いた対話シーンにおける注目度推定
Image based Attention Estimation for Interaction Scene

小宮 凛子† 齊藤 剛史† 嶋田 和孝†
Rinko Komiya Takeshi Saitoh Kazutaka Shimada

1. はじめに

対話は，雑談や教育現場での討論，職場での会議など様々な場面で行われている．対話シーンの解析はコミュニケーション能力改善や教育現場での学習効果の評価に役立つ．発言者に関しては，音声情報を用いた発話内容の解析や映像情報を用いたジェスチャの解析などがある．一方，非発言者は，頷きや顔の向き，視線など映像情報を解析することで，対話への集中度などの内面状態の推定や理解度を判断することが可能となる[1]．本研究では対話シーンの映像情報から，参加者の内面状態の推移を測ることを目的とし，本稿では，非発言者が注目している人物を自動的に推定する手法を提案する．

石井らは，複数人対話における発言者交替のパターンを予測しており，注視情報を考慮することで発言開始タイミングの予測性能が向上することを示している[2]．Ba と Odobez は頭部姿勢を用いて会議参加者の注意集中推定に取り組んでいる[3]．Otsuka らは，全方位カメラを用いてグループ会議のリアルタイム分析システムを提案しており，視線方向や発言者を自動的に推定している[4]．しかし，いずれも複数の撮影機材を必要としており，解析対象となる顔の画像サイズは大きい．本稿では全天球カメラ 1 台を用いて撮影した複数人対話シーンに対して提案手法を適用し，その有効性を確認する．

2. 提案手法

2.1 集中度

本稿では，非発言者が発言者を注目する度合を注目度と定義する．注目度には，個人単位の注目度 A_i と対話全体の注目度 A_{all} の二つが考えられる．前者は，発言者 P_i の発話フレーム数 N_i に対する非発言者 P_j が P_i を注目しているフレーム数 $N_{i,j}$ の比率 $A_i = N_{i,j} / N_i$，後者は A_i の平均値と定義する．

注目度を推定するために，発言者の発話フレームおよび非発言者が注目している人物を求める必要がある．発話フレームに関しては，本稿では手動で与えるが，例えば音声情報を利用したり，口元の動きを観察することで発話フレームを求めることが期待される．ここでは，非発言者が注目している人物を自動的に推定する手法を提案する．

一般的に参加者が注目している人物は，参加者の頭部姿勢および視線から求めることができるが，本稿では頭部姿勢のみを用いて注目人物を推定する．人間の頭部姿勢は，pitch，yaw，roll の 3 次元で表現される．本研究では参加者は環状に配置された椅子に着席した状態における対話シーンを対象とする．このため，頭部姿勢は yaw，すなわち水平方向のみを考慮すればよい．参加者 P_j の頭部水平方向角度 θ_j と P_j に対する他の参加者 P_i の位置方向 $\alpha_{j,i}$ の角度差 $d_{j,i} = |\theta_j - \alpha_{j,i}|$ を求め，条件 $d_{j,i} < T$ を満たす場合に，P_j は P_i を注目していると判定する．T は注目しているか否かを判定する角度しきい値である．

2.2 頭部角度推定

本研究では図 1 に示すような全天球カメラ 1 台で撮影された画像に対して equirectangular 形式でパノラマ画像に変換した対話シーン画像を入力とする．

まず，画像中から HOG ベースの検出器を用いて参加者の顔を自動的に検出する．検出された顔領域に対して，顔部位の特徴点を検出し，著者らが提案した特徴点ベースの頭部姿勢推定手法[5]を適用して，参加者の頭部水平方向角度 θ を推定する．

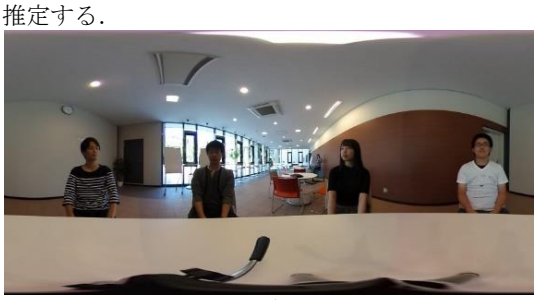

図 1　対話シーン

図 1 のシーンは円卓を囲み参加者 4 人が椅子に座っているシーンである．カメラから各参加者までの距離は同じである．このシーンに対して，検出された顔位置および推定角度を図 2 のように可視化する．本稿ではこの図を対話マップ画像と呼ぶ．中央の円は円卓を表現しており，円卓周辺にある四つの小さな円は，顔検出結果より推定された参加者位置である．また各参加者位置より伸びる太い赤線は推定角度 θ を表している．参加者円の右上に表記されている青色数字は参加者 ID，右下に表記されている赤色数字は推定角度である． このマップ画像より，参加者 1，2，3 は参加者 0 を注目していることが確認できる．

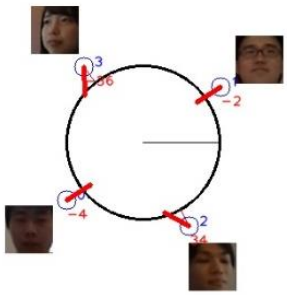

図 2　対話マップ画像

†九州工業大学，Kyushu Institute of Technology

2.3 参加者の注目人物の決定

前節により参加者 P_i の頭部角度θ_iが得られている．また全参加者の位置が得られているため，P_i に対する他の参加者 P_j の位置方向$\alpha_{j,i}$も算出できる．これにより，θ_i と$\alpha_{j,i}$ の角度差$d_{j,i} = |\theta_j - \alpha_{j,i}|$が求まる．対話シーンにおける参加者数を M とすると，P_i からは参加者自身を除く$M-1$の角度差が計算される．そこで条件$d_{j,i} < T$を満たす i に対して，$i^* = \mathrm{argmin}_i\, d_{j,i}$より P_iの注目人物 P_{i^*}を決定する．

発話フレームに対して各参加者の注目人物を求め，注目度を推定する．

3. 評価実験

3.1 対話シーン

本研究では参加者数が 3〜5 人で構成される複数人対話シーンを撮影した．撮影には RICHO 社製 THETA S を用いた．equirectangular 形式のパノラマ画像変換には同社が提供するパソコン用アプリケーション RICOH THETA を用いた．変換されたパノラマ画像の画像サイズは 1920×960 画素であり，パノラマ画像における参加者の顔領域サイズは約 50×50 画素であった．撮影時のフレームレートは 30fps であった．シーン数は 7 であり，参加者数 3 人が 2 シーン，4 人が 2 シーン，5 人が 3 シーンである．参加者は円卓を囲むように椅子に着席した．

提案手法を定量的に評価するためには，対話シーンにおける各フレームの発言者，および，各参加者がどの参加者を注目しているかの正解情報が必要である．定量的に評価しやすいように，対話シーンのタスクを設計し，参加者はこのタスクに従って対話させた．具体的には，参加者は 1 人ずつ順番に 1 回ずつ発言し，発言中は，発言者は正面を向き，非発言者は発言者の方向に頭部を向けるように指示した．ただし，推定結果が発言者の順番に依存しないように，対話シーンごとに発言者の順番を変更した．これにより，評価対象フレームは，発言しているフレームのみであり，このフレームは音声情報を用いて手動で抽出した．また提案手法により，非発言者は発言者方向を注目していると推定されれば成功，そうでなければ推定失敗と判断でき，容易に定量的に評価可能となる．

頭部角度推定のための顔検出および特徴点検出には，機械学習ライブラリ Dlib C++ Library [6]を用いた．

3.2 実験結果

7 シーンに対して提案する注目度推定法の精度を評価した．角度しきい値を T = 10 度とした場合の全シーンの平均推定精度は72.7%であった．参加者数毎にわけた場合，3 人シーン，4 人シーン，5 人シーンの平均推定精度はそれぞれ 100%，95.9%，40.0%であった．参加者数が 3，4 人の場合，高い精度で推定されているが，参加者数が 5 人の場合は精度が低い．この原因を調査するために，シーン毎の精度でなく，非発言者の注目方向に対する注目度推定精度を算出した．その結果をプロットしたものを図 3 に示す．横軸は非発言者の正面を 0 度とした場合の注目方向であり，縦軸は注目度推定精度である．注目方向の絶対値が大きい場合，推定精度が低いことが確認できる．これは提案手法で用いる頭部姿勢推定は特徴点ベースであり，角度が大きくなり横を向くと特徴点検出精度が低下し，頭部姿勢推定に誤差が含まれるためである．この結果から，参加者数が少ない場合，参加者同士の距離が離れることで，非発言者は頭部方向を大きく変えなくても発言者を注目することが可能であるため，高い注目度推定精度を得ることができているが，発言者が多い場合（5 人）には，頭部方向が大きく変わり，精度が大幅に下がってしまったものと考えられる．

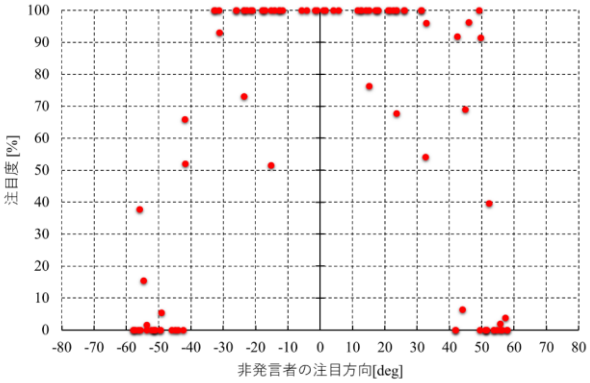

図 3　注目方向に対する注目度推定結果

4. おわりに

本稿では，対話シーンから対話参加者の頭部姿勢を用いて，非発言者の注目度を推定する手法を提案した．撮影した対話シーンを用いた評価実験の結果，注目人物からの発言者の位置角度が小さい，すなわち正面に近い位置に発言者がいる場合は高い推定精度があることを確認した．

注目度推定の精度向上には，頭部姿勢推定の精度向上が必須である．今後の課題として頭部姿勢推定の改善が挙げられる．また本稿では，自然な対話シーンでなく，設計された対話シーンを解析対象とした．自然な対話シーンに対して提案手法を適用し，評価を検証することも課題である．

謝辞

本研究の一部は，JSPS 科研費 17H01840 の助成によるものである．

参考文献

[1] 横川 和章, 有馬 道久, “教授場面における非言語的コミュニケーション：理解状態の表出と判断”, 教育心理学研究, Vol.34, No.2, pp.120-129 (1986)

[2] 石井 亮, 大塚 和弘, 熊野 史朗, 松田 昌史, 大和 淳司, “複数人対話における注視遷移パターンに基づく次話者と発話開始タイミングの予測”, 信学論, Vol.J97-A, No.6 (2014).

[3] Ba, O. S. and Odobez, J.-M., “A Study on Visual Focus of Attention Recognition from Head Pose in a Meeting Room”, In *Proc. of MLMI2006*, pp.75-87 (2006).

[4] Otsuka, K., Araki, S., Ishizuka, K., Fujimoto, M., Heinrich, M., and Yamato, J., “A Realtime Multimodal System for Analyzing Group Meetings by Combining Face Pose Tracking and Speaker Diarization”, In *Proc. of ICMI*, pp. 257-264 (2008)

[5] Komiya, R., Saitoh, T., Fuyuno, M., Yamashita, Y., and Nakajima, Y., “Head Pose Estimation and Motion Analysis of Public Speaking Videos”, *International Journal of Software Innovation*, Vol.5, Issue.1 pp.57-71 (2017).

[6] DLib C++ Library, http://dlib.net/.