視聴覚情報を統合した話者注意対象の検出

陳　彬　　目黒　光彦　　金子　正秀

電気通信大学　大学院電気通信学研究科　電子工学専攻

1.　はじめに

ユーザと人間共存型ロボットとの間での柔軟なインタラクションを行うためには、ユーザの注意対象の検出が要素技術として重要である。ロボットの注意先をユーザの注意対象に向かせることは「共同注意の形成」と呼ばれている[1]。筆者らは、これまでに、顔姿勢の推定に基づいてロボットとユーザとの間で共同注意を形成する方法について研究を行ってきた[2]。本論文では、視聴覚情報を統合して、複数ユーザによる会話シーンの中から話者の定位を行い、ロボットと話者との間での共同注意を交替的に形成する方法について述べる。

2.　ハードウェア構成

図１に、人間とのコミュニケーションに必要な視聴覚センサーを備えた全方向移動可能なロボットを示す。ロボットの頭部に立体視可能な３眼カメラを取り付けている。さらに、5 個のマイクロホンにより構成されたマイクロホンアレーを実装している。３眼カメラと各マイクロホンとの位置関係は既知である。

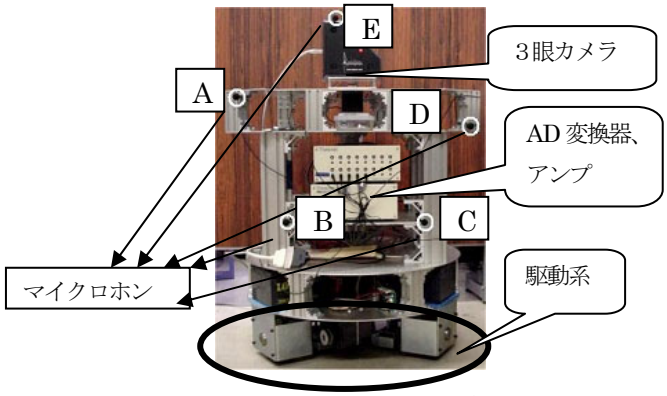


図 1　ハードウェア構成

3.　３次元音源定位の原理

従来の３次元音源定位方法では、広い領域に配置した大規模なマイクロホンアレーが必要になる。本論文では、移動ロボットでの利用という観点から、画像と音声情報を統合することにより、大規模なマイクロホンアレーを必要としない３次元音源定位方法を提案する。

- Detection of Speaker's Focus of Attention by Integrating Auditory and Visual Information
- Bin Chen, Mitsuhiko Meguro, and Masahide Kaneko
- The University of Electro-Communications

５個のマイクロホンにより構成された 10 組のマイクロホンペアの各々について、音声信号に対する白色化相互相関係数(CSP)を計算する。３眼カメラと各マイクロホンとの位置関係が既知であるので、３眼カメラで計測した実空間中の点から発せられた音声信号の各マイクロホンペアへの到来時間差を計算することができる。この時間差における CSP 係数を、実空間中の点に対応する画像平面内の座標に書き込み、音源位置情報を表す２次元的な尤度マップを生成する。各マイクロホンペアに対して得られたマップを加算すると、和のマップにおいて、真の音源位置に対応した場所に尤度の最大値が得られる。そこで、尤度の最大値を求めることにより、音源を３次元的に定位することができる。

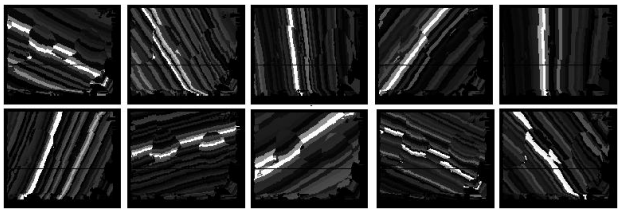


図 2　各マイクロホンペアにおける音源定位の尤度マップ

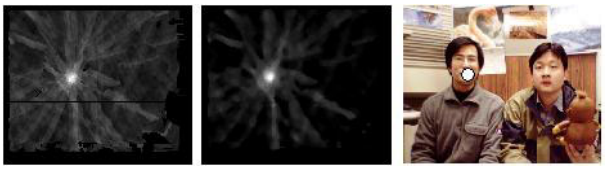


図 3　音源定位の結果. (a) 各マップを加算した結果. (b) (a)を正規化したマップ. (c) 音源定位結果.

図１に示したマイクロホン A, B, C, D, E によって構成された 10 組のペアに対して、音源定位の尤度マップを計算した結果を図２に示す。図２の左上から順にペア AB, AC, AD, AE, BC(以上、上段), BD, BE, CD, CE, DE(以上、下段)で得られた結果を示す。図中の明るい領域は高い CSP 係数を持ち、音源がそこに存在する可能性が高い。図２の各マップを加算した結果を図 3(a)に示す。 (a)のマップに対して、平均値以下の値を 0 で置き換えて平滑化を行った結果を(b)に示す。さらに、音源位

置を画像上へマッピングすることにより話者の抽出が行える。この結果を(c)に白い円で示す。

4. 共同注意形成の原理

顔姿勢の情報は比較的得やすいため、ユーザの注意対象の検出への利用が期待できる。そこで、本論文では、顔姿勢に基づいてユーザの注意対象の検出を行う。

まず、シーンのテクスチャ画像と距離画像を用い、人物の頭部の 3 次元ワイヤフレームモデルを利用して、顔の特徴点の 2 次元画像平面上での動きから、最小二乗法に基づいて実空間中での 3 次元的動きを推定する。

次に、ユーザの視野を楕円錐体で近似する。楕円錐体の軸をユーザの視線に対応させ、人間の網膜上の生理的視力分布の特性を定式化する。実空間内の点に対して、図 4(b) に示す距離画像とユーザの顔姿勢情報を利用して、ユーザの網膜における視力相対値(感度値を 0 ~ 1 に正規化した値) を計算することができ、画像と同じサイズを持つ「視力分布マップ」と呼ばれる 2 次元のマップが得られる(図 4(c))。

(a) (b) (c)

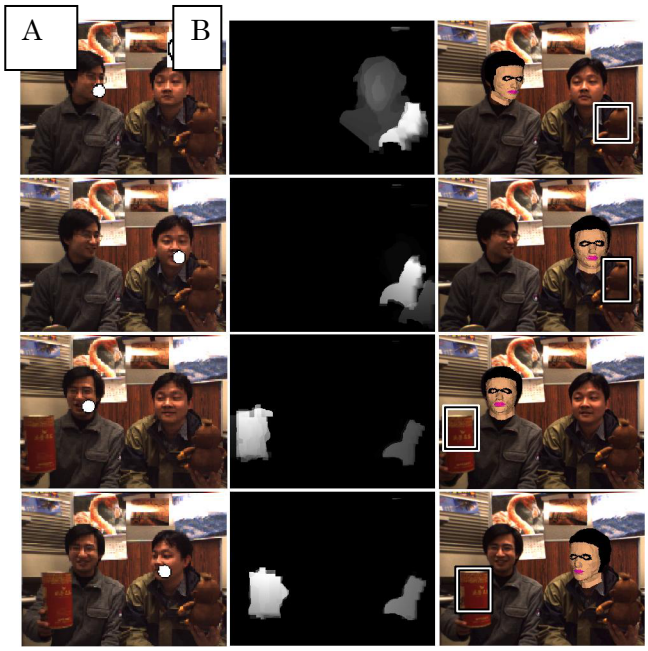
図 4 (a) 入力画像. (b) 距離画像. (c) 視力分布マップ.

最後に、視力分布マップで正の値を持つ画像領域に対して動きを求める。動物体領域における視力分布を、領域の動きに従ってフレーム間で伝搬させ、時間累積和を計算する。ここで得られた累積和のマップは、ユーザの目の前の異なる位置に配置された物体の各々がロボットに対して異なる誘目度を持つことを意味するので、ユーザとロボットとの「共同注意形成の誘目度マップ」と呼ぶことにする。さらに、誘目度マップを確率分布マップと見なして、CamShift アルゴリズム[3]により、物体領域を同定する。

5. 交替会話シーンにおける話者との共同注意の形成

複数話者による会話シーンにおいて、話者定位と話者の顔追跡を併用して、話者の交替を検出する。ロボットの注意先を、検出した話者の注意対象に向かせることによって、ロボットと話者との共同注意の形成を行う。

図 5 において、向かって左のユーザを A さん、右のユーザを B さんとする。(a)は音源定位の結果であり、白い円が検出された話者(口の位置)を表している。(b)は誘目度マップ、(c)は話者の注意対象の検出結果(四角い枠で囲った物体)である。最上段の第 42 フレームにおいて、ロボットは話者が A さんであると判断して、A さんの注意対象である「ゴジラのぬいぐるみ」を抽出した。第 2 段の第 91 フレームでは、話者が B さんに移り、B さんの注意対象も「ゴジラのぬいぐるみ」と抽出された。第 3 段の第 115 フレームでは、話者が再び A さんに代わり、注意対象が「お茶の缶」であることが分った。以降のフレームにおいても、このような処理が続けて行われている。図 5 の結果より、ロボットと話者との共同注意が交替に形成されていることが分かる。



(a) (b) (c)

図 5 交替会話シーンにおける話者との共同注意の形成. (a) 音源定位結果. (b) 共同注意形成における誘目度マップ. (c) 話者注意対象の検出結果.

6. むすび

本論文では、複数の参加者による会話状況を理解するために、話者とロボットとの間での共同注意を交替的に形成する方法について検討した。今後、より複雑なシーンへの適用、実時間処理などの課題について研究を進めていきたい。

参考文献

1) B. Scassellati : “Foundations for a theory of mind for a humanoid robot,” MIT Department of Computer Science and Electrical Engineering, PhD Thesis, 2001.

2) 陳 彬, 目黒 光彦, 金子 正秀 : “顔姿勢推定に基づくユーザとロボットとの共同注意の形成,” 平成 14 年電気学会, 電子・情報・システム部門大会講演論文集, OS5-9, 2002.9.

3) G. R. Bradski : “Computer vision face tracking for use in a perceptual user interface,” Intel Technology Journal, Q2, 1998.