

EM アルゴリズムとパーティクルフィルタの統合によるリアルタイム複数の人物追跡システム実現

金鉉燦[†] 駒谷和範[†] 尾形哲也[†] 奥乃博[†]

京都大学大学院 情報学研究科 知能情報学専攻[†]

1. はじめに

従来の視聴覚情報統合による人物追跡システムには特定方向から来るノイズや全方向の追跡に制限があったり[1]、マイクロホンアレイおよび測定したインパルス応答が必要だった[2]。しかし、実用的な音源定位や人物追跡には、事前情報を用いず、入力されたデータのみからリアルタイムに複数追跡するシステムが必要である。人間の耳と同じ数である二本のマイクロホンを用いれば装置の設置制限や費用が削減できる。しかし、二本のマイクロホンの CSP は、多音源がある環境における音源定位結果のあいまい性および誤差が増加するので実用的ではなかった。本研究では EM アルゴリズムを用いて入力されたデータの分布を推定することによって複数の音源定位結果のあいまい性を解消し、信頼度を高めた。そして、その結果にパーティクルフィルタを適用して複数人物の追跡を行った。

2. システム構造

図 1 の左側は人物追跡に使用したロボット SIG2 である。両側の耳のところに二本のマイクロホンが装着されており、二個のカメラが設置されている（現在は一つのカメラのみ使用）。

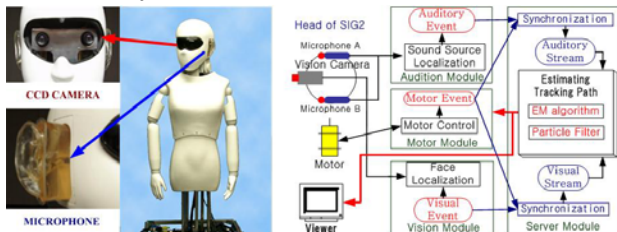


図 1: SIG2 ロボットとシステムの概要

図 1 の右側はシステムの構造である。全部で五つのモジュールで構成されており、それぞれの機能は以下の通りである。

音声処理 (Audition) : 約 32ms ごとに CSP (Cross-power Spectrum Phase) による音源定位を行い、その情報を人物追跡モジュールに送る。

画像処理 (Vision) : 約 250ms ごとに顔検出および顔定位を行い、その情報を人物追跡モジュールに送る。顔検出には商業用ライブラリーの OpenCV を用いた。

人物追跡 (Server) : 音声や画像処理モジュールから受信したイベントデータを利用し、EM アルゴリズムとパーティクルフィルタを用いて人物追跡を行う。

モーター (Motor) : 人物追跡モジュールから受信した人物追跡結果にしたがってモーターを回転させる。SIG2 は頭の回転範囲が-90 度から 90 度まで、腰は-45 度から 45 度

までなので-135 度から 135 度まで人物追跡できる。

ビューアー (Viewer) : 各モジュールの出力結果および追跡情報を画面に表示する。

3. 人物追跡

複数の人物追跡や前方位の追跡を行うためにはマイクロホンアレイや全方位カメラを使えば高精度な人物追跡ができる。しかし、それらのセンサーを使用する場合は費用が高く、ロボットに設置するための物理的な制約が大きい。したがって本研究では二本のマイクロホンでインパルス応答が必要ない CSP を用いて音源定位を行った。ただし、二本のマイクロホンの CSP によって二個以上の音源定位を行う場合、誤差が増加して音源定位性能が低下する。その短所を解決するため、提案手法として一定区間 (300ms) の音源定位データに EM アルゴリズムを適用してデータ分布推定を行うことにより、データのあいまい性を解消し、複数の音源定位を実現した。本手法では複数のセンサーからの定位情報を統合することができ、本稿では音源定位と顔定位の統合を行った。そして動いている複数の対象を同時に追跡する場合、EM アルゴリズムによって検出された各人物定位結果にパーティクルフィルタを適用して複数の人物追跡を行う。

3.1 複数の音源定位実現および視聴覚情報統合

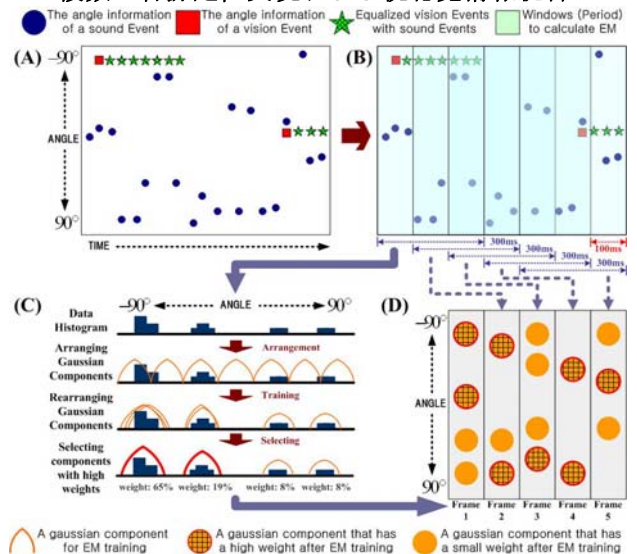


図 2: EM アルゴリズムによるデータ分布推定

本節では複数の音源定位および視聴覚情報を統合するために EM アルゴリズムを用いて、データ分布推定を行う方法について説明する。データ分布推定の概要を図 2 に示す。(A) は音源定位 (丸) および顔定位 (四角) でのそれぞれの角度情報を表す。ここで音源定位のイベント発生間隔 (約 32ms) と顔定位のイベント発生間隔 (約 250ms) が

異なるのでサンプリングレートを統一するために顔定位データを補完する(図の星)。そして(B)では 300ms 区間の窓を掛け、EM 学習させると窓内のデータの分布が推定される。角度軸の全領域にガウス分布を配置して EM 学習させると(C)のようにヒストグラムの密度が高い領域に収束する。そこで重なっていたガウス分布は一つに合わせ、重みの大きい分布は音源が存在する位置と判断する。そして同じ方法で窓を 100ms ずつずらしながら EM 学習させ、推定された結果を(D)のようにフレームに順次保存する。

3.2 複数の追跡経路推定

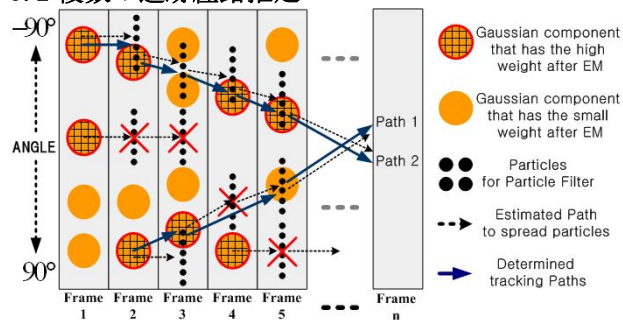


図 3：パーティクルフィルタによる複数経路推定

追跡方法は図 3 のように EM の結果から重みが高いガウス分布(格子縞)の中で追跡に使われなかったものからパーティクルフィルタを適用して追跡を始める。追跡が始まったら過去経路の位置、速度、加速度を考慮して次の追跡位置を予測し、その位置の±15 度領域にパーティクルを振りかけて正確な経路を求める(Path 2)。そして追跡中の場合は一定区間の間にガウス分布がなくても持続的に追跡が可能である(Path 1)。一定区間以上、予測した位置にガウス分布が存在しなければ追跡をしない。

4. 評価実験

本手法の性能を評価するために二つ実験を行った。

1. 二個の音源がある場合、各音源に対してどれくらい追跡ができるのかについて実験した(図 4)。
2. 実環境での顔定位情報統合による人物追跡をした(図 5)。

結果はビューアーに表示された追跡結果を示した。図の中で緑点と青点はそれぞれ発生された音源定位と顔定位情報を示し、赤線が提案した手法で推定された追跡経路である。そしてピンクの領域はカメラの視野であり、真ん中は頭モーターの現在回転された角度を示す。

追跡性能を評価するために、音源間の距離および移動速度を変えて音源追跡を行った。ロボットから 1.5m 離れたところに二個のスピーカーをおいて各々男性と女性音声を持続的に 85dB の大きさで出力した。図 4 の左側は EM とパーティクルフィルタを適用した結果、右側は EM のみを適用した結果である。(A)と(B)はロボットの頭を 1rpm (Revolutions Per Minute)速度で回転させた結果であり、(C)と(D)は 3rpm で行った結果である。そして(A)と(C)のスピーカー間隔は 30 度である。(B)と(D)は 60 度である。実験結果より、本手法を適用した場合、音源間距離が 30 度以上、音源速度が 3rpm 以下ならば正しく追跡できることを確認した。

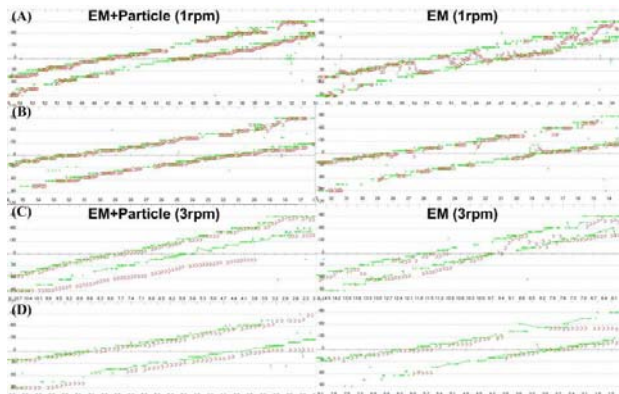


図 4：二音源の追跡結果

顔定位統合による人物追跡ではロボットが先に検出された追跡対象または現在方向から近い追跡対象に頭を回転する。また実環境ではテレビやラジオまたは音楽などの音声が混ざったノイズが多いと仮定して、実験では一定の方向に音源が持続的に検出されればロボットの頭(カメラ)をその方向に回転させる。そして一定時間以上(3 秒ぐらい)顔検出が無い場合はノイズと判断し、そちら側の方向に対して音源定位による追跡は行わないようにする。追跡結果では、同時に二個のスピーカーから再生される音声を追跡し(A)、先に検出された 45 度方向に頭を回転した。ノイズと判断した後、追跡を止めた(B)。次に-30 度に回転し、同じ方法でノイズと判断して追跡を止めてから戻った(C)。そして-120 度から出た人物の声に反応し、回転してから顔定位結果と統合して 120 度まで追跡した。実験結果より、音源定位と顔定位を統合した本手法によって、複数音源の中で移動しながら発話する人物を正しく追跡できることが分かった。

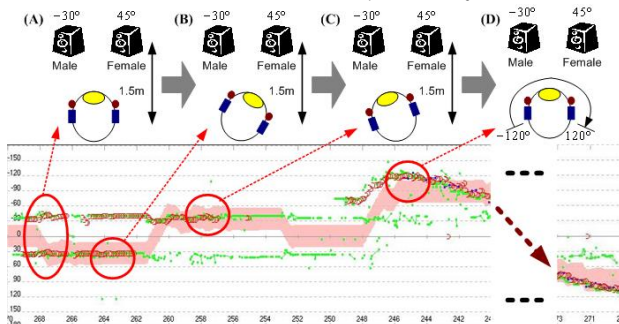


図 5：ノイズがある環境での人物追跡結果

5. まとめ

本稿では、EM アルゴリズムを用いて音源定位結果と顔定位結果を統合し、複数音源定位結果のあいまい性を解消した。そして、その結果にパーティクルフィルタを適用して複数の人物追跡を実現した。複数音源の追跡実験により、本手法の有効性を確認した。

参考文献

- [1] K. Nakadai *et al.*, "Real-Time Auditory and Visual Multiple-Object Tracking for Humanoids," IJCAI-01, pp. 1425-1432.
- [2] I. Hara *et al.*, "Robust Speech Interface Based on Audio and Video Information Fusion for Humanoid HRP-2," IROS-2004, pp. 2404-2410.