I-036

ベイジアンネットによる視聴覚情報の統合を利用した会話シーンからの話者検出

Speaker Detection in Conversational Scene Based on Audiovisual Integration with Bayesian Net

陳 彬 * 目黒 光彦 * 金子 正秀 *
Bin Chen Mitsuhiko Meguro Masahide Kaneko

1. はじめに

人間を相手にしているのと同じような感覚でユーザがロボットとインタラクションを行うためには, 画像, 音声などの視聴覚情報を利用することが有用である. 複数のユーザによる会話シーンでは, 視聴覚情報を分析して話者を特定すると共に, 話者の 3 次元空間位置を推定することが必要となる.

3 次元音源位置の推定に関しては, マイクロホンアレーを用いる方法がある. しかし, マイクロホンアレーを利用して推定した音源位置は, 一般的に空間解像度が低く, マイクロホンアレー座標系における音源位置を画像平面の座標系に変換して, 画面から音源領域を正確に検出することは困難である. この問題を克服するために, 視聴覚情報の統合により, 画像から話者を直接検出して追跡する研究が行われている [1, 2]. しかし, これらの研究においては, 話者の画像特徴の追跡や唇などの動き情報の検出を必要とするため, 背景領域の動きが単純であること, 話者の唇の画像解像度が確保されることを前提としている. また, シーンの 3 次元情報を利用していないため, 話者の 3 次元位置情報を得ることは難しい. これに対して本論文では, ベイジアンネットを利用して視聴覚情報を統合することによって, 3 次元的な話者位置を画像から適切に検出する方法を提案する.

2. 多チャネル音信号と距離画像との統合処理

本論文で用いたロボットには, 頭部に立体視可能な 3 眼カメラが取り付けられており, 環境の奥行き距離を測定することが可能である. ロボット本体の正面には 5 チャンネルのマイクロホンアレーが実装されている. 3 眼カメラと各マイクロホンとの間の座標関係は既知である.

カメラからの可視音源を対象とし, 3 眼カメラを利用して音源の候補位置の集合を距離画像により予め獲得しておく. マイクロホンペアを A, B とした場合, 距離画像上の各画素に対応する空間中の点を仮想音源とし, 各仮想音源から発せられた音の信号がマイクロホン A, B へ到達する時間の差を幾何学的に求めることが出来る. この時間差におけるマイクロホン A, B での音の信号の白色化相関 (CSP) 係数を求め, 音源方向を表す 2 次元的な尤度マップを生成する. 以上の処理を各マイクロホンペアに対して施すと, 本論文の場合, 10 個のマップが生成される. 各マップについて累積和を計算すると, 音源位置を推定することが可能な尤度マップが得られる [3]. この尤度マップを $R(i,j)$ により表す. ただし, (i,j) は画面座標である. $R(i,j)$ から最大値を求めることにより, 音源位置情報が得られる.

*電気通信大学　大学院電気通信学研究科電子工学専攻

3. SVM による音信号の分類

一般に, 音声音の音源位置と画像中の肌色特徴との間で関連性が高い. 一方, 非音声音の音源位置と肌色以外の特徴との関連性が高くなる. このため, 音信号を音声音/非音声音へ分類しておくことが有用である. 音信号の分類は, ガウスカーネルを持つサポートベクターマシン (SVM) を用いて行う. 本論文では, ATR 音声データベースと RWCP 非音声音データベースを学習データとして利用する. 音信号に対して, 現時刻から過去の 1 秒間内に発生した音の信号をサンプルとし, ① Spectrum Flux, ② LPC 係数, ③ Band Width, ④ Band Centroid, ⑤ Band Energy Ratio の計 5 種類の特徴を抽出し, サポートベクターマシンを学習させる.

4. GMM による肌色分布のモデル化

人種や照明条件が多様である場合, 画像内の画素を肌色/非肌色へ分類するためには, ガウス混合モデル (GMM) を適用することが有効である [4]. 本論文では, 異なる人種の人物について多様な照明条件下で収集した肌色サンプルに対して, HSV 空間における H と S 成分の分布を 4 つのガウシアンを有する GMM によりモデル化する.

5. ベイジアンネットによる視聴覚情報の統合

ベイジアンネットは, 事象をノード, 各事象間の因果関係をリンクにより記述した有向グラフである. ネットワークの構造を定めた後に, 学習サンプルを用いて, ネットワークパラメータであるノードの条件付き確率表 (CPT) を学習する. 証拠となるノードの値を観測した場合, ネットワークの条件付き確率表に基づいて, 推論の目的となるノードでの値の確率分布を推定することができる.

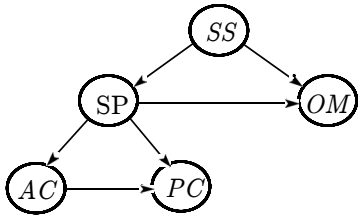

Fig.1: Bayesian net for speaker detection.

本論文で用いるベイジアンネットの構造を Fig.1 に示す. 画面内の各画素に対して, このベイジアンネットを利用して, 「ある画素が話者と対応するかどうか」という意味を持つノード SP に対して推論を行う. Fig.1 では, 各ノード間の依存関係を次のように考える. 画面上のある画素 (i,j) が, 話者と対応していれば $(SP=1)$, SP の親ノードが $(SS=1)$ となる. SS が音源を表し, 1 と 0 の値を持つ. それぞれ真と偽を表す. この時に, R マップ上

の同一位置にある画素から高い尤度値 $R(i,j)$ を出力する．これをノード OM により記述する．ノード SP, OM の各々は 1 と 0 の値を持つ．また，音声信号が発生した場合 ($SP=1$)，音の種類に対する認識結果 ($AC=1$) を出力する．ノード AC には 0 と 1 の値があり，非音声音と音声音を表す．さらに，音信号の認識結果 (AC) 及び SP の値により，その画素における画像特徴が肌色/非肌色であることがわかる．これを PC で示す．PC にも 0 と 1 の値があり，それぞれ非肌色，肌色特徴と対応する．ベイジアンネットの構造を定めた後，最尤推定法を用いて CPT を計算する．

2. と **4.** によって，ノード OM, PC の値が 1 を取る時の尤度が得られる．また，**3.** での SVM からの出力は音声音である場合，尤度 $Pr_{(i,j)}(AC=1)=1$ とする．すなわち，$Pr_{(i,j)}(OM=1)$, $Pr_{(i,j)}(PC=1)$, $Pr_{(i,j)}(AC=1)$ をソフト的証拠として，ベイジアンネットを用いて推論を行い，尤度 $Pr_{(i,j)}(SP=1)$ を推定することができる．$Pr_{(i,j)}(SP=1)$ を「信念度マップ」と名付ける．信念度マップから見つけ出した最大値が，ある閾値以上であれば，話者が最大値の座標に存在する可能性が高い．さらに，画像から話者を検出した後に，距離画像から話者の 3 次元位置を得ることができる． また，尤度マップ $Pr_{(i,j)}(SS=1)$ 上の最大値を求めることにより，音声音/非音声音の音源位置を推定することもできる．

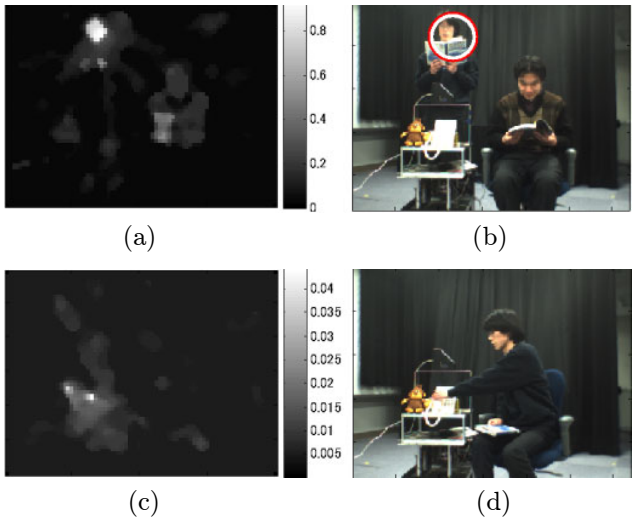

Fig.2: (a) Belief map of referencing at the 61st frame. (b) The speaker marked with ◯ is detected according to (a). (c) Belief map at the 145th frame. (d) According to (c), no speakers are detected due to low belief values, even though there is an audio signal generated when the user put down the receiver of telephone.

Fig.2(a), (c) は，学習したベイジアンネットを第 61, 145 フレームの画面上の各画素に適用して得られた信念度マップを示す．後方の話者が本を読み上げている第 61 フレームに対して，信念度マップ (a) から見つけ出した最大値が 0.931 であり，最大値の座標上に話者が存在する可能性が高い．この結果を (b) に ◯ により示す．一方，第 145 フレームは，受話器が置かれる時に，電話機本体との物理的接触により音信号を発生するシーンである．(c) は，このような非音声音を含んだシーンに対して得られた信念度マップ $Pr_{(i,j)}(SP=1)$ である，信念度マップ (c) から見つけ出した最大値は 0.042 であり，シーン内に話者が存在する可能性は低い．このため，(d) には ◯ が表示されていない．Fig.3 の (a)~(f) では，フレーム 34, 45, 72, 97, 101, 170 においての話者検出結果を，図中の話者に ◯ を重畳して示している．話者が正しく検出されていることがわかる．

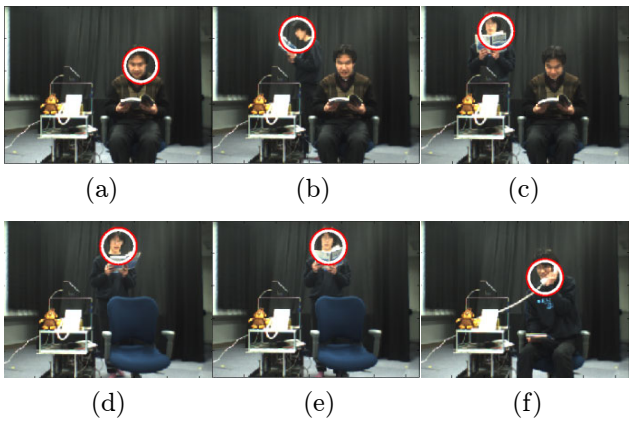

Fig.3: Results of speaker detection. (a)~(f), Results of speaker detection obtained at the 34th, 45th, 72nd, 97th, 101st, 170th frame, respectively.

6. むすび

本論文では，視聴覚情報の統合により，画像から話者を検出する方法を提案した．本論文での提案手法は，ユーザとロボットがより円滑なコミュニケーションを行うためのインタフェース，人間型ロボットの視覚システムなどに適用することができる．今後，ロボットの周囲にある物体から発せられた音信号に対しての認識結果と，音源に対して視覚的注意を向けた後に取得した画像的特徴との結び付けを自動的に行い，ベイジアンネットにリンクやノードの値を追加し，話者のみならず，非音声音の音源位置も安定して推定する方法を検討していきたい．

参考文献

[1] M.J. Beal, N. Jojic, and H. Attias: “A graphical model for audiovisual object tracking,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.25, no.7, pp.828-836 (July 2003)

[2] J.M. Rehg, K.P. Murphy, and P.W. Fieguth: “Vision-based speaker detection using Bayesian networks,” Proc. of Computer Vision and Pattern Recognition, vol.2, pp. 110-116 (June 23-25, 1999)

[3] 陳彬, 目黒光彦, 金子正秀: “会話シーンにおけるロボットと複数ユーザとの共同注意の形成,” 映像情報メディア学会誌, vol.57, no.7, pp.854-863 (July 2003)

[4] Quan Huynh-Thu, Mitsuhiko Meguro, and Masahide Kaneko: “Skin-color extraction in images with complex background and varying illumination,” Proc. of IEEE WACV2002, pp.280-285 (Dec. 2002)