I-39　全方位カメラと複数のマイクロホンを用いた話者の検出

Detection of Talking People using Omni-directional Camera and Microphones

土田 勝† 川西 隆仁† 村瀬 洋† 高木 茂†
Masaru Tsuchida Takahito Kawanishi Hiroshi Murase Shigeru Takagi

1．まえがき

様々な通信技術の発展に伴い、大量の情報の流通や蓄積が盛んに行われてきている。多様な情報を処理し、状況に応じて有用な情報を我々に提示する一つの形態に、擬人化エージェントの活用がある。例えばイベント会場などで、情報をリアルタイムに更新し、来場者の要望に合わせて提示することができれば、より効率的な活動が可能となる。また遠隔地の人々との対話においても、エージェントが中継することでより潤滑なものになりうる。ここで重要となるのが、話し手（話者）の位置の特定である。現在の人物とエージェントとのコミュニケーションは、1対1の形態をとることが多い。一方、複数の人物やエージェントが混在する環境においては、エージェントが話者の方を向き、話しかけたり表情を変えたりすることが重要で、そのことでより親密なコミュニケーションが実現できる。しかしながら、複数対複数の対話環境における話者の位置検出は、画像、音声等のみを用いたユニモーダルなシステムでは十分な精度を得ることは困難であり、様々な入力情報を効果的に組み合わせたマルチモーダルなシステムの構築が必要となる。

本論文では、画像情報と音声情報を用いて話者の方向を特定する手法について述べる。今回、全方位カメラ、2つのマイクロホンおよび円筒型表示装置[1]を組み合わせたシステムを作成した。本システムでは、全方位カメラで撮影した画像を用いて室内に存在する複数の人物位置の追跡を行い、さらにその結果とマイクロホンアレイからの音響信号とを用いて話者を検出し、その方向にエージェントが顔を向け、対話を始めることができる。2章ではCondensationアルゴリズム[2]を基にした人物の追跡、および、話者を検出する手法について説明し、3章では上述のシステムで行った実験結果について述べる。

2．複数の人物が存在する環境での話者の検出

Condensationアルゴリズム[2]はモンテカルロ法に基づく追跡手法である。対象物の状態(位置)を確率密度で表現し、常に複数の解の候補を持ちながらの追跡が可能であり、またノイズに対するロバスト性も強い。本節では、Condensationアルゴリズムによる追跡手法について簡単に説明した後、全方位画像における人物の追跡手法と、マイクロホンによる受音信号を用いた話者の検出について述べる。本論文においては、二人以上が同時に声を出すことはないと仮定する。

2.1　Condensationアルゴリズム

時刻tにおける追跡対象の状態を X_t、画像から得られる観測結果を Z_tとする。また、時刻tまでの状態と観測結果をそれぞれ $\mathbf{X}_t=[X_1, \cdots, X_t]$, $\mathbf{Z}_t=[Z_1, \cdots, Z_t]$とする。このとき追跡問題は、時刻tにおいて状態が X_tとなる確率密度 $p(X_t|\mathbf{Z}_t)$を推定する問題と考えることができる。ここでベイズの法則により次式が成り立つ。

$$p(X_t|\mathbf{Z}_t)=k_t\, p(Z_t|X_t)\, p(X_t|\mathbf{Z}_{t-1}) \qquad (1)$$

$$\text{ここで、} p(X_t|\mathbf{Z}_{t-1})=\int p(X_t|X_{t-1})p(X_{t-1}|\mathbf{Z}_{t-1})dX_{t-1}$$

あらかじめ時刻 t-1 での推移確率 $p(X_t|X_{t-1})$が与えられている場合には、各時刻における尤度 $p(Z_t|X_t)$を画像から推定することで $p(X_t|\mathbf{Z}_t)$を得ることができる。実時間で全ての状態 X_tに対して尤度を求めることは非現実的であり、通常はランダムサンプリングの結果を用いて推定を行う。

2.2　全方位画像における人物の追跡

全方位カメラを中心とした空間において、どの方向に人物が存在するかを検出し、またその動きを追跡する。全方位画像には、画像の中心からの放射線上に、それぞれの方向に対応する情報が記録されている。そこでこれらの情報を画像から抽出して処理を行う。本論文では予め撮影しておいた背景画像を用いて、画像中の人物と背景を分離し、得られた差分画像を用いて人物の追跡を行う。

はじめに、差分画像に対し二値化処理を行う。そして得られた画像に対しCondensationアルゴリズムを適用し、人物の追跡を行う。まず、二値化した画像から各方向に相当する成分を抽出し、その方向での人物の存在に対する尤度を推定する。ここでは抽出した画像成分に含まれる人物の領域の面積に注目し、時刻tにおける θ_k方向に対する尤度 $\pi^{(k)}_t$を次式のように定義する（図1）。

$$\pi^{(k)}_t=(\Sigma I^{(k)}_t)/S\ , \qquad (2)$$

$$\theta_k=k\cdot 2\pi/N,\ \ N:\text{サンプル数}$$

$I^{(k)}_t$とSは抽出した画像領域の画素値と面積である。

時刻 t-1 におけるサンプルと尤度を $s^{(k)}_{t-1}$、$\pi^{(k)}_{t-1}$ [k=1,…, N] と表わす。まず、サンプル $s^{(k)}_{t-1}$を尤度 $\pi^{(k)}_{t-1}$の比に従って選択し、推移確率に従ってランダムに変化させる。これ

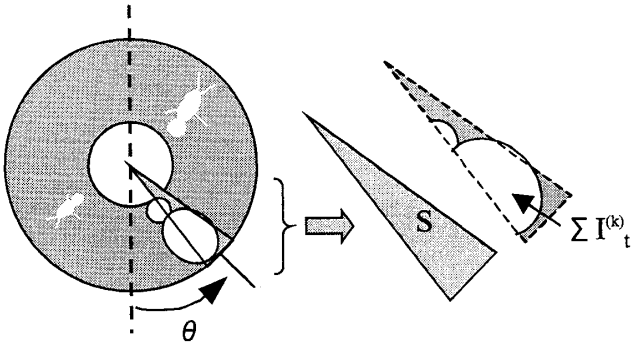


図1: 全方位画像からの尤度の推定
左の図の白い部分は、人物をあらわしている。

† 日本電信電話株式会社　NTTコミュニケーション科学基礎研究所

をN回繰り返し、時刻tにおけるN個のサンプルを新たに生成する。さらに次の時刻 t+1 におけるサンプルは、時刻tのサンプル $s^{(k)}_t$ と尤度 $\pi^{(k)}_t$ から求める。

2.3 話者の検出

話者は $\pi^{(k)}_t=0$ となる方向を除いた、いずれかの方向に存在すると考えられる。そこで尤度 $\pi^{(k)}_{t-1}$ の比に注目し、その値が極端に低いサンプルはノイズとみなし排除する。その後、残りの全てのサンプルについて、それぞれの方向に音源があった場合に、2つのマイクロホンで受音される音響信号間の時間遅れを計算する。マイクロホン同士の間隔をd、音速をcで表わすと、k番目のサンプル(方向 θ_k)における受音信号間の時間遅れ τ_i は次式を用いて求めることができる。

$$\tau_k = (d \sin \theta_k) / c \qquad (3)$$

計算結果を基に、実際に受音した音響信号の一方に τ_k の時間遅延を行い、2つの音響信号間の相関を調べる。同様の処理を全てのサンプルについて行い、相関値が最大となる方向に、話者がいるとみなす。

3. 実験

原理的な確認を行うため、全方位カメラおよび2本のマイクロホンを用いて実験を行った。図2に実験で用いたシステムの外観を示す。全方位カメラは凸型の双曲面ミラーとCCDカメラで構成され、XGAの画像が取得できる。マイクロホンの間隔は25cmとし、カメラと共に円筒型表示装置の上部に取り付けた。円筒型表示装置にはエージェントを表示し、推定結果に合わせて話者の方向を向く。以下では、室内にいる2人のうち1人が、エージェントから見て右45度の方向から話しかけた際の実験結果を示す。

図3(a)に全方位画像、(b)に背景と人物を分離した結果の画像を示す。また図4には、図3(b)から求めた各方向に関する尤度と、音響信号から求めた相関値（共に最大値で規格化）を示す。なお今回の実験では、全方位画像における人物追跡の際にランダムサンプリングは行わず、2度間隔でサンプリングした180方向に関して尤度を求めた。こより話者の方向は右48度となり、良好な推定結果が得られた。図5に推定された方向を画像上で示す。

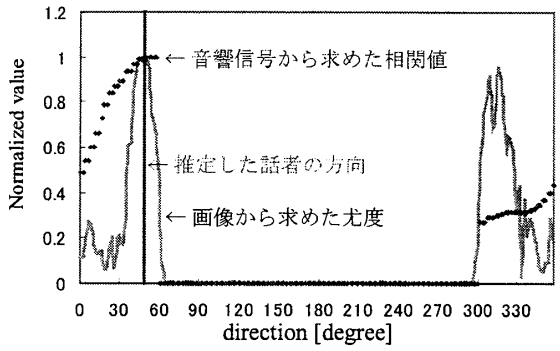


図4: 話者がいる方向の推定

曲線は画像から求めた尤度を、点線は音響信号から求めら相関値を示す。縦線は話者の方向を推定した結果（右48度）を示す。

4. まとめ

全方位カメラと2つのマイクロホンを組み合わせ、室内に複数の人物が存在する環境において話者の方向を特定する手法を提案した。また、円筒型表示装置と組み合わせたシステムを構築し、提案手法の有効性を確認した。今回はマイクロホンを2本しか用いなかったため、エージェントの前方の人物に関する検出のみを行った。今後は使用するマイクロホンの本数を増やし、提案手法の有効性に関し更に検証を進めていく予定である。

参考文献

[1] 川西、土田、村瀬、高木 "知的エージェントの表示を目的とした小型円筒ディスプレイの提案." 信学技報 *PRMU02*, Jul., 2002.

[2] Micheal Isard and Andrew Blake. "Condensation - condentional density propagation for visual tracking." *International Journal of Conputer Vision*, Vol. 29, No. 1, pp.5 28, 1998.

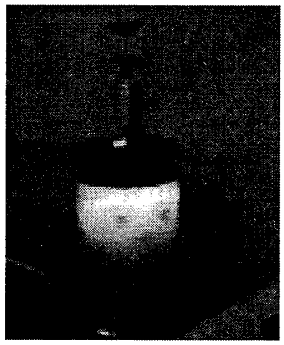
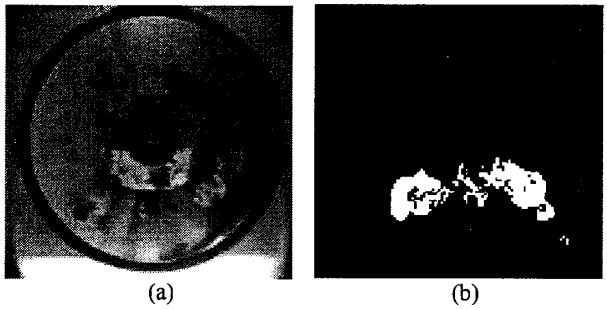


図2: 実験で用いたシステムの外観



(a) (b)

図3: 全方位画像

2人の人物が写っている。(a) 入力画像,
(b) 背景と人物を分離して二値化を行った結果

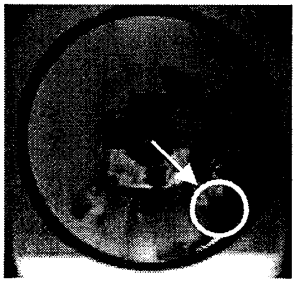


図5: 話者の検出結果