論　文

相互情報量最大化に基づく信号情報源の移動軌跡の推定

池田　徹志†　　石黒　浩†　　浅田　稔†

Moving Signal-Source Tracking Based on Mutual Information Maximization

Tetsushi IKEDA†, Hiroshi ISHIGURO†, and Minoru ASADA†

あらまし　従来の異種センサの信号統合手法は，各センサで個別に特徴抽出処理を行った後に統合を行うアプローチが主流である．このアプローチの問題点として，特徴抽出による抽象化によって失われた情報が統合時には利用できないことが挙げられる．本論文では，統計的手法に基づき信号間の統合を早期に行う新しい統合手法を提案する．統計的手法を用いた信号統合では，観測対象の移動等による動的な環境への適応が問題となる．これに対し信号間の相互情報量を最大化する規範に基づき，対象の追跡を行いながら信号を統合する手法を提案する．音と画像の信号統合を例として，カメラとマイクロホン各１台を用いた移動音源の位置推定の実験を行い，提案手法の有効性を確認した．

キーワード　音源位置推定，信号統合，センサフュージョン，相互情報量，知覚情報基盤

1. はじめに

複数の異種センサの情報を統合して用いるセンサ統合の研究が精力的に行われている [16] [17]．異種のセンサを組み合わせることの利点は，様々なモダリティを通して対象を多様な側面からとらえた表現を得られることである．一方で異種のセンサで観測された信号は一般に信号の性質も大きく異なるため，統合のための信号変換が必要になる場合が多い．従来の異種センサ統合のアプローチでは，各センサの信号に対して特徴抽出処理を独立に行い，その後に得られた結果を統合するアプローチが主流であった．

信号統合の新しいフレームワーク

これに対し人間の知覚系では，異なるセンサの信号をより早い段階から統合を行っていると考えられる．例えば，音声と音声を発する口の動きの視聴覚的な統合に関しては多くの知見が得られている．視覚が音声の認識に大きな影響を与える McGurk 効果 [11] をはじめ，視覚情報を用いることにより音声の認識精度が音声のみのときと比べて高まることが知られており，視覚が音の知覚の基本的な段階で影響を与えていることが示唆されている [7]．また音声言語以外でも，視覚，聴覚，触覚などさまざまな感覚が，認識過程で互いに大きな影響を与え合うことが知られている [14]．各感覚の認識処理による抽象化が完了する以前の段階から統合を行うことにより，抽象化によって失われる情報を統合の時点で利用できると考えられる．工学的な異種センサ統合においてもこのアプローチに注目することにより，より高度なセンサ統合処理の実現が期待できる [3]．

従来の工学的な異種センサの信号統合のアプローチでは，各センサに対して設計者が目的とするタスクに適切と考えた特徴を抽出し，その後に特徴量を統合する手法が一般的である (図 1)．このアプローチの問題点は，複数のセンサで同一の対象を観測した場合に信号そのものに表れている信号間の変化の同期性や絡み合いなどの情報が特徴抽出によって失われ，統合の段階で利用できない点である．

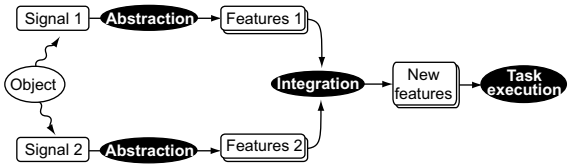

図 1　従来のセンサ統合のアプローチ
Fig. 1　Previous approaches of sensor integration.

† 大阪大学大学院工学研究科，大阪府
Graduate School of Engineering, Osaka University, Yamadaoka 2–1, Suita, Osaka, 565–0871 Japan

これに対し，信号間の変化の同期性などの情報を積極的に利用し，特徴抽出処理を行う以前のできるだけ早期の段階で信号を統合するアプローチが考えられる(図 2)．本論文の目的は，この早期の信号統合のアプローチの可能性を示すことである．

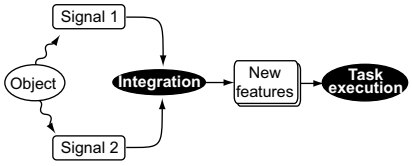

図 2　提案する統合のアプローチ
Fig. 2　The proposed approach of sensor integration.

関連研究

近年，信号の統計的性質を利用し，センサの統合を早期の段階で行う手法が提案されている．Becker [1] [2] は，画像を入力とする２つのニューラルネットワークを，出力間の相互情報量を最大化するという規範で学習することにより，入力画像間の視差に相当する特徴を抽出できることを示した．

異なる種類のセンサ信号を統合する研究では，Hershey et al. [8] は，交互に話す２人の話者をカメラとマイクロホン１台で観測した場合に，画像中の話者の領域の画素は音声と同期して変化することに注目した．音声と画素値の時系列の相互情報量を求め，高い値を示した画素を選択することによって話者の位置を画像上で推定できることを示した．Fisher III et al. [6] は，信号のモデルを仮定せずに信号間の相互情報量を計算する方法を提案し，情報量を最大にする斜影変換を求めることにより，話者の位置を画像上で推定し，また指定された画像領域に存在する話者の音声を強調するフィルタリングを行った．Nock et al. [12] は画像中の話者の位置を検出する問題に対し，音声と画素値の時系列の関係を評価する各種の方法を実験的に比較し，同時ガウス分布を仮定して相互情報量を計算する手法が最も良い結果になることを示した．Kidron et al. [9] は音に対応して変化する画素は局在すると仮定し，パラメータ数に対してデータが少ない場合でもロバストに推定する方法を提案した．

これらの手法の特徴は，信号間の統計量を用いることにより，性質が全く異なる信号の統合を直接的に行うことができる点である．しかし，上記の研究では各センサが同一の対象を観測し続けていること，すなわち観測信号間の統計的性質が定常であることが仮定されており，統合処理が固定的なことが本質的な問題であった．そのため，音源位置の推定の際には信号源が移動しないことが仮定されており，環境の中で信号源が移動してゆくような現象に適用することができなかった．

音源が画面上で移動する場合にも適用できる手法では，Slaney and Covell [15] は画像中から顔領域の抽出を行い，音声の特徴ベクトルと画像の輝度値との間の正準相関分析に基づき，画像中から音声信号と同期して変化する画素を検出する方法を提案した．Li et al. [10] は音声と画像の信号に対し，信号が最も類似するような斜影変換を行った後に，両者の相互情報量を求める手法を提案した．また，Fisher III and Darrel [5], Nock et al. [13] では抽出された顔領域に対し，前述の同一著者の手法 [6] [12] が適用されている．

しかしこれらの研究では，顔領域の抽出処理後に，抽出された領域に対して同様の手法を適用する手法であり，音源対象の画像認識用のモデルが必要になり適用対象が限られる．さらに本質的な問題点として，統合処理の前に認識処理を行う手法であり，認識処理に失敗した場合には統合処理も失敗するためロバストではない．

この問題に対して本論文では，異種のセンサの信号に対して認識処理を行う前の段階において，相互情報量を最大化する最適化による信号統合の枠組みを提案し，信号源の移動などのセンサの信号間の統計的性質の変化に対して適応的に対応できることを示す．はじめに，相互情報量を対象の移動した軌跡に沿って計算する手法を提案する．これにより信号源が移動した場合でも，信号間の関係を正しく抽出することができる．次に，信号間の相互情報量を最大化する規範に従い，対象の位置推定と移動する信号源に適応した信号間の関係の抽出を同時に行う手法を提案する．提案手法の有効性を確認するため，カメラとマイクロホンを用いた音源位置推定を例として，移動音源の位置推定を行う実験を示す．

以下，2. で相互情報量を用いた信号統合について述べ，従来手法の問題点を指摘し，相互情報量最大化を規範とする信号統合手法を提案する．3. では提案する統合手法を，カメラとマイクロホンを用いた音源位置推定を例として具体的に述べる．4. で実験結果を報告し，5. でまとめを述べる．

2. 相互情報量の最大化に基づく異種センサの信号の統合

2.1 相互情報量を指標とした共通情報源の検出に基づく信号統合

同一の情報源を複数の異種センサで観測した信号は，モダリティの違いによる表層的な表現の相違はあるが本質的には同じ情報を含み，一方が変化すれば他方も同期して変化するなど互いに関連して変化する．よって複数の信号の一方を観測した時に他方の値を良く予測できるならば，両者は信号源を共有する信号ではないかと考えることができる．一般に，ある確率変数の値によって他の確率変数の値について知りうる情報の量は変数間の相互情報量として定義される．従って複数の信号が情報源を共有するかどうかを，一方から他方を予測するときの情報量に注目し，相互情報量を指標として評価することが有効と考えられる．

信号間の相互情報量を計算することにより，対象の認識処理など抽象化を行う以前の早期の段階で，異種センサを統合した情報源の検出を行うことができる．アレイ状にセンサを配置したカメラなどのセンサと，別のセンサとを用いて情報源を観測するとき，センサアレイ中の各センサと後者のセンサとの間で信号間の相互情報量を計算する．求められた相互情報量が大きいセンサを選択することにより，共通情報源の位置をセンサアレイ上で検出できる．以上は Hershey et al. [8] によって提案された手法であり，近年研究が盛んな統計的性質を利用した異種センサ統合のアプローチの基本的なアイディアである．

2.2 従来手法の問題点

文献 [8] で提案された統合手法では，センサの信号間の相互情報量などの統計量を計算する間は信号間の統計的性質が定常であることを仮定している．従ってセンサアレイ上を信号源が移動して観測される場合などの動的な場合には適用できないという問題点がある．

この問題点を，カメラとマイクロホンを用いて音源を観測する例で説明する．従来の信号統合手法に基づいて相互情報量を計算する様子を模式的に図 3 に示す．信号源が環境中で一定位置に存在する場合，各画素と音信号との間の相互情報量を求めることにより，音源の位置を推定することができることを表している．

しかしセンサアレイ上を信号源が移動すると，信号源を観測するセンサは時間と共に入れ替わる．観測を開始した時点で音と特定の画素の間に現れた信号の相関関係は，信号源が移動するに従って他の異なる画素との間で観測されるようになる．したがって，長時間の観測によって音と特定の画素の間の統計的関係を求めても，信号源が発する信号の間の関係を捉えることはできない (図 4)．

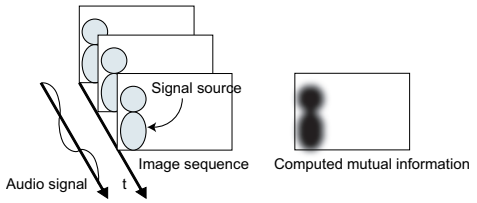

図 3　信号源が静止している場合における相互情報量計算
Fig. 3　Computing mutual informaiton between sensory signals when the signal source does not move.

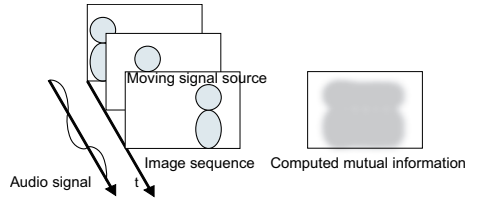

図 4　信号源が移動する場合における相互情報量計算
Fig. 4　Computing mutual informaiton between sensory signals when the signal source moves.

2.3 移動軌跡に沿った相互情報量の計算

この問題に対し，センサアレイ上で信号源を最も良く観測するようなセンサを常に切り替えて得られる信号を用いることにより，信号源の移動に適応する手法を提案する．画像と音の信号統合の例では，画像上の信号源の移動軌跡上に存在する画素の輝度値の時系列と音声信号との間の相互情報量を計算することにより，信号源が発する信号間の関係を捉えることができると考えられる (図 5)．

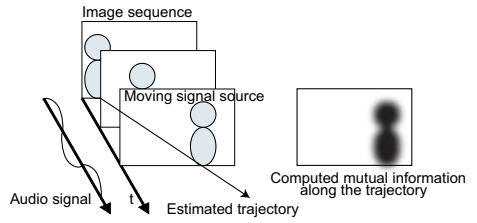

図 5　対象の軌跡に沿った相互情報量の計算
Fig. 5　Computing mutual information along the trajectory of the target.

2.4 相互情報量最大化に基づく移動軌跡の推定

しかし信号源の軌跡は一般に未知である．これに対し，軌跡の候補の中から信号間の相互情報量を最大にする軌跡を求めることにより，信号源の移動軌跡を推定する手法を提案する．複数のセンサで観測した信号間の相互情報量を最大にする最適化問題を解くことにより信号統合を行い，統一的に信号源の軌跡の推定を行う．信号源の位置は軌跡に沿って計算した相互情報量を評価関数として，評価関数を最大にする軌跡を探索的に求めることによって得られる．この様子を図 6 に示す．

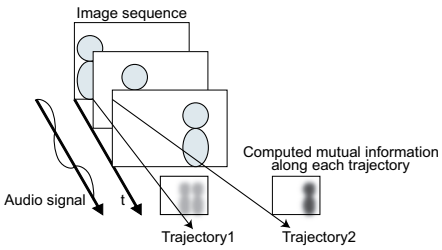

図 6　相互情報量最大化の規範による対象の軌跡の探索
Fig. 6　Detection and tracking a target based on mutual information maximization between sensory signals.

2.5 モデルの導入

信号源の軌跡推定を探索的に行う際に，信号間の相互情報量は弱いヒューリスティクスであるため，特に探索の初期段階では良い評価関数とならず，探索が幅優先的になってしまう．効率よく確実に正解軌跡を求めるため，以下を導入する．

領域を単位とした相互情報量の評価

センサアレイ上の各要素センサの移動軌跡を個別に推定した場合，正解とは異なる軌跡に沿って計算した相互情報量が偶然大きくなる場合を排除することが難しい．環境内で静止した背景部分を観測した要素センサの信号であっても，特定の移動軌跡に対しては相互情報量がゼロにはならず，ノイズとして現れる．

そこで本研究では，各要素センサごとに軌跡を評価するのではなく，センサの領域を単位として軌跡を評価することを提案する．領域内の要素センサは平行に移動すると考え，各要素センサの軌跡に沿って相互情報量を計算し，その平均値を評価値とする．

画像領域を単位とした軌跡推定を行うことにより，多くの要素センサが平均的に相互情報量を大きくする軌跡を求めることができるため，ノイズの影響を受けにくくなる．領域のサイズは扱う対象に応じて選択する方法や，最も良く対象の信号間の関係を捉える最適なサイズを適応的に求めるなどの方法が考えられる．ここでは固定値を用いた．

信号源の移動軌跡のモデルの導入

任意の移動軌跡の集合から相互情報量を最大にする軌跡を求める問題では，探索空間が膨大になり，現実的に解を求めることが困難である．本論文では信号源の移動軌跡のモデルを導入し，信号源の移動に現実的な制約を加えることにより探索空間を絞り込む．移動モデルは扱う対象に合わせて適宜選択するものとし，ここでは信号源の速度に上限を設定して速度を区分的に一定とする単純な移動モデルを導入する．

3. 音と画像の直接統合による音源の定位

提案する統合手法を，環境中に設置されたカメラとマイクロホンを用いて音源の位置推定を行うタスクを例として，具体的に説明する．

3.1 音と画像間の相互情報量の計算

音および画像中の画素の時系列信号をそれぞれ $A = A(t), V = V(t)$ とすると，$A(t)$ と $V(t)$ の間の相互情報量は以下のように求められる．

$$\mathrm{I}(A;V) = \mathrm{H}(A) + \mathrm{H}(V) - \mathrm{H}(A,V) \tag{1}$$

ここで $\mathrm{H}(A)$，$\mathrm{H}(V)$ はそれぞれ $A(t), V(t)$ のエントロピ，$\mathrm{H}(A,V)$ は両者の結合エントロピである．

$$\mathrm{H}(A) = -\sum_t \mathrm{p}(A(t)) \log \mathrm{p}(A(t)) \tag{2}$$

$$\mathrm{H}(V) = -\sum_t \mathrm{p}(V(t)) \log \mathrm{p}(V(t)) \tag{3}$$

$$\mathrm{H}(A,V) = -\sum_t \mathrm{p}(A(t), V(t)) \log \mathrm{p}(A(t), V(t)) \tag{4}$$

ここでは $A(t)$ と $V(t)$ が同時ガウス分布に従うと仮定し，相互情報量を近似的に以下のように求める ([8], 導出の詳細は [4]).

$$\frac{1}{2}\log\frac{1}{1-\rho(A,V)^2} \tag{5}$$

ここで $\rho(A,V)$ は $A(t)$ と $V(t)$ の相関係数である．

同時ガウス分布を仮定した相互情報量の近似計算に関する議論

式 (1) の計算の目的は，音声と画像の信号が情報源を共有するかどうかを評価するため信号間の相互情報量を求めることである．ここでは相互情報量を実際に評価する際に近似計算を行う方針について述べる．

A(t),V(t) のエントロピおよび結合エントロピを推定する方法として，信号の確率分布が未知の場合には，信号の確率分布のモデルに仮定を置かない方法 ([6] など) が一般性の高い方法である．しかし相互情報量を十分な精度で推定するためには比較的多くのデータ量が必要になるため，長時間の観測が必要になり，信号源や観測系が動的に変化する場合に厳密な計算が困難になる場合がある．

このような場合には信号の分布の複雑さに応じ，適切なモデルを選択して近似的に計算することが有効である．式 (5) では A(t),V(t) が同時ガウス分布に従うと仮定して相互情報量を近似的に求めている．A(t),V(t) の分散共分散にのみ注目して推定することになるため厳密性を欠くが，特に信号間の関係が定常ではないなどデータ量が限られる場合には有効と考えられる．文献 [12] では音声と画像の信号レベルの統合において各種指標を実験的に比較検討した結果，同時ガウス分布近似に基づく相互情報量が最も優れていたと報告されている．

しかし同時ガウス分布の近似が有効でないことが判明した信号に対しては，モデルを仮定しない方法の適用や，同時ガウス分布よりも複雑なモデルを適用する方法や，得られるデータの量に応じてモデルを切り換える方法などについて，現実的な適用可能性を検討する必要があり，今後の課題である．

3.2 画像上の対象の軌跡に沿った相互情報量の計算

時刻 t における画像上の対象の位置を $x(t)$ とすると，一般に式 (3)(4) は以下のように $x(t)$ を含んだ式となる．

$$\mathrm{H}(V) = -\sum_t \mathrm{p}(V(t,x(t)))\ \log\ \mathrm{p}(V(t,x(t))) \tag{6}$$

$$\mathrm{H}(A,V) = -\sum_t \mathrm{p}(A(t),V(t,x(t))) \log\ \mathrm{p}(A(t),V(t,x(t))) \tag{7}$$

3.3 音と画像間の相互情報量最大化に基づく音源位置推定

音源位置推定の問題を，信号間の相互情報量を最大化する軌跡を求める最適化問題と考え，以下に示すアルゴリズムにしたがって探索的に解を求める．対象の画像上での移動軌跡を仮説とし，多数の仮説を評価値の高い順に一定数保持するビームサーチにより探索を行う．仮説の評価関数として移動軌跡に沿った音と画像の相互情報量を用いる．

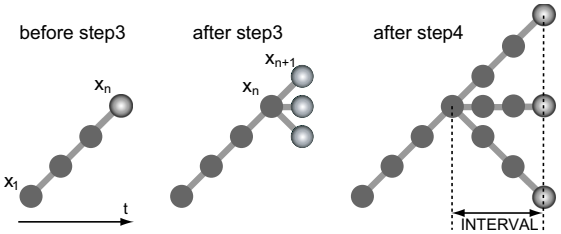

図 7 探索時の仮説の展開．仮説スタック中の各仮説が探索木の葉に対応する．

Fig. 7 Expansion of hypotheses. Each hypothesis in the hypothesis stack corresponds to a leaf of the search tree.

アルゴリズム

（1） 対象の存在する可能性のある各位置 x について，長さ 1 の軌跡を持つ仮説 h_x を生成する．これらの仮説の集合を仮説スタックの初期値とする．

（2） 仮説スタック中の仮説の持つ軌跡の長さが T に達した場合，仮説スタックから相互情報量の最も高い仮説を取り出し，探索を終了する．

（3） 仮説スタック中の各仮説 h について，仮説 h が持つ対象の軌跡を $\{x_1 \ldots x_n\}$ とすると，対象の次時刻での予測位置 x_{n+1} を接続した新たな仮説 h' を生成する (一般に予測の不確定性に対応して複数の仮説が生成される)．予測は導入した移動軌跡のモデルに従って行う．仮説スタックから h を除き，生成された新たな仮説を加える．

（4） 仮説スタック中の各仮説 h について，対象の次時刻での予測位置を接続した新たな仮説 h' を生成する．ここで予測位置は，導入した区分的に速度一定の移動モデルに従う．仮説スタックから h を除き，生成された新たな仮説を加える．このステップを INTERVAL - 1 回だけ繰り返す．

（5） 仮説スタック中の各仮説 h について，軌跡に沿った音と画像間の相互情報量を計算する．

（6） 仮説スタックを (5) で計算した相互情報量の大きい順にソートし，評価値の順に最大 BEAM 個のみを残す．

（7） (2) に戻る．

ここで T，INTERVAL，BEAM は定数とする．BEAM は保持する仮説数の上限であり，BEAM を十分大きくすることで最適解が求まるが，探索に必要な計算量は大きくなる．また，音源の移動のモデルとして，最大速度に上限を設定し，INTERVAL フレームの間は速度一定とするモデルを導入している．(3) での現在の軌跡の終端位置 x_n と予測位置 x_{n+1} の差が速度に相当し，(4) の処理は INTERVAL フレームの間，その速度に固定して仮説を展開していることに相当する．

4. 実　　験

提案手法の有効性を検証するため，音信号と画像信号を統合し，画像上で移動する音源の位置を推定する実験を行った．部屋に２人がいる環境で，１人は足音を立てて移動し，もう１人は音を立てず手を振っている様子を，カメラとマイクロホン各１台を用いて観測した．

画像の解像度は 180×120 であり，30 フレーム/秒で取得した．音は 16kHz でサンプリングし，画像の１フレームに対応する 1/30 秒ごとの平均パワーを求めたものを用いた．以降は時間の単位を画像の１フレームとする．音と画像の信号の例を図 8，図 9 に示す．図中の右の人が足音を立てて歩き，左の人が手を振っている．

図 8　実験に用いた画像例
Fig. 8　An example image.

4.1 従来手法による音源位置推定

従来の信号レベルの統合手法 (式 (2)-(4)) に基づき，画像の各画素の値と音の平均パワーとの間の相互情報量を求めた結果を図 10 に示す．相互情報量の大きさを濃度で表してあり，濃い領域は相互情報量が大きいことを示す．

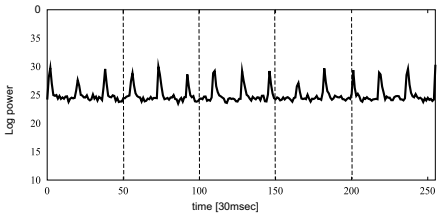

図 9　実験に用いた音信号
Fig. 9　An example of sound signal.

音源の移動に伴い，計算された相互情報量は音源位置に集中せずに画像上で広がっているため，結果から音源の位置を推定するのは困難である．

図 10　従来の直接統合手法による結果
Fig. 10　Sound source localization result with previous direct fusion method.

4.2 相互情報量最大化に基づく音源位置推定

提案手法に基づき，音声と画像間の相互情報量を最大化する規範に基づいて探索的に音源位置の推定を行った．探索のパラメータは以下のように設定した．今回の実験では，移動モデルは水平方向の移動のみとし，探索も図の水平方向にのみ行っている．

パラメータ	値
BEAM	1000
INTERVAL [frame]	32
速度の上限 [pixel/frame]	0.1

探索を右の歩行者の領域から開始した時に，探索によって軌跡の推定を進める様子を図 11 の左側のグラフに示した．グラフの横軸はフレーム単位の時間，縦軸が画像中の水平方向の位置を表す．開始から 64, 128, 192, 256 フレーム (それぞれ，約 2,4,6,8 秒) 経過した時点での，評価値が第５位までの仮説に対応する軌跡を図中の太線で，第 10 位までを点線で示した．図の右側に，対応する時点の評価値最大の軌跡に沿って計算した領域内の相互情報量の大きさを濃度で表し，相互情報量の平均値を右肩に数値で示した．探索が進む

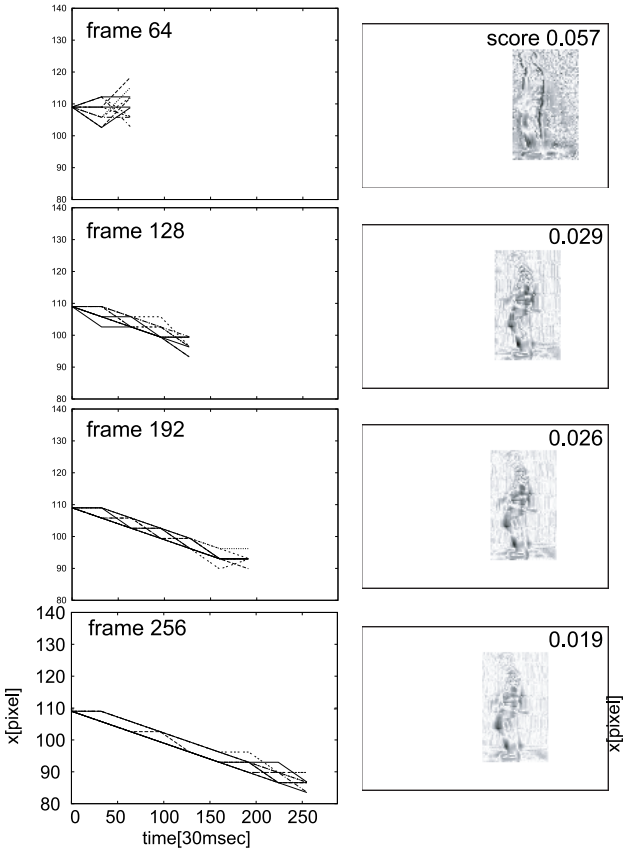

図 11 右の人の領域から探索を開始した場合の t=64,128,192,256[frame] における探索の経過．各時点での上位 5 位までの仮説を実線で，10 位までの仮説を点線で示す．

Fig. 11 The five best hypotheses (solid line) and ten best hypotheses (dashed line) at t=64,128,192,256 frame. Initial position of the hypothesis is in the right person.

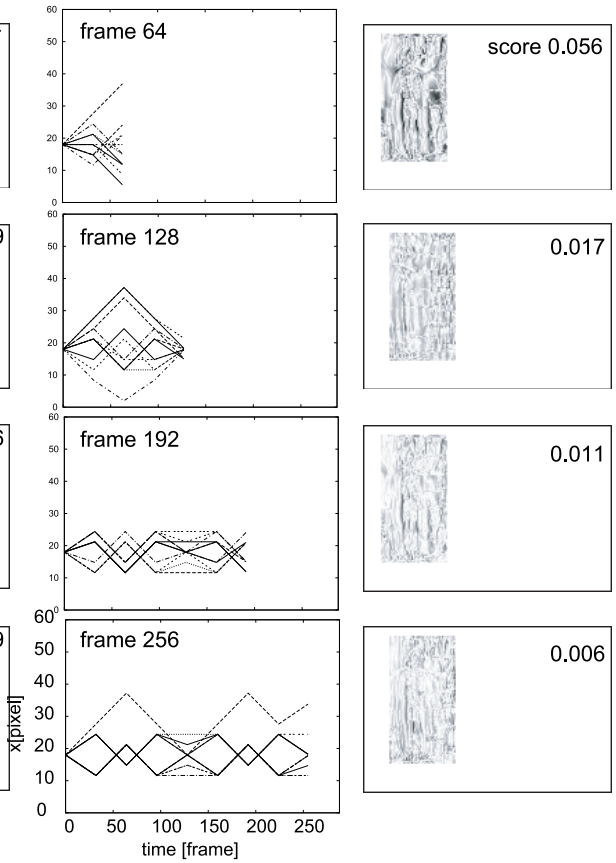

図 12 左の人の領域から探索を開始した場合の t=64,128,192,256[frame] における探索の様子．各時点での上位 5 位までの仮説を実線で，10 位までの仮説を点線で示す．

Fig. 12 The five best hypotheses (solid line) and ten best hypotheses (dashed line) at t=64,128,192,256 frame. Initial position of the hypothesis is in the left person.

にしたがって足音の音源である足付近の領域の相互情報量が相対的に大きくなり，音源位置の位置が絞り込まれてゆく様子が分かる．

同様に，探索を左の手を振る人の領域から開始した場合の結果を図 12 に示す．こちらの領域では音信号と関係のある動きは観測されないため，探索が進むにつれて相互情報量は急速に小さくなってゆく様子が分かる．

4.3 信号波形の比較

図 13，図 14 に求められた軌跡上の画素値と，音信号の信号の様子を抽出したものを示す．図 13 は手を振る人の領域上の 1 点 (手付近) での画素値を，図 14 は歩行者の領域の 1 点 (足付近) での画素値の変化を示している．音信号は比較のため，図 9 と同じものを並べて示した．

図 13 では信号間に同期した変化は見られないのに対し，図 14 ではピークが同期して現れていることが分かる．

5. ま と め

信号間の相互情報量を最大化する規範に基づき，異種センサの信号をより早期に統合する新しい信号統合のフレームワークを提案した．はじめに観測するセンサが切り替わる移動信号源に対し，対象の軌跡に沿って相互情報量を計算する手法を提案した．次に信号間

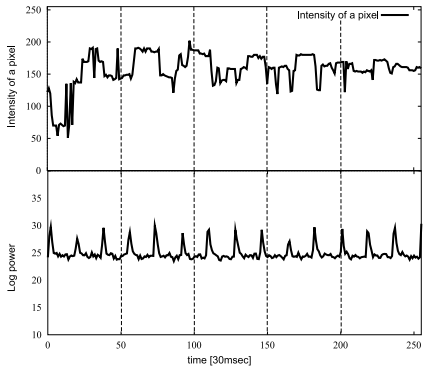

図 13　画素値 (上) と音 (下) の信号例 (手を振る人)
Fig. 13　Sequence of average power of sound signal (upper) and intensity of a pixel in the region of the left person (who shakes his hand, lower).

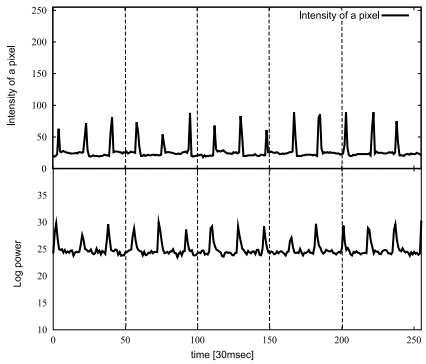

図 14　画素値 (上) と音 (下) の信号例 (歩く人)
Fig. 14　Sequence of average power of sound signal (upper) and intensity of a pixel in the region of the right person (who walks, lower).

の相互情報量を最大化する規範に従い，対象の移動軌跡を求めながら信号の統合を行う手法を提案した．

音と画像の統合を例として，カメラとマイクロホン各１台を用い，画像上で移動する音源の位置を推定する実験を示した．複数の人が動いている様子が画像上に観測されている場合でも，足音を手がかりにして歩行者の位置を推定することができることを示し，提案した信号統合のフレームワークの可能性を示した．

今後の課題として，信号がガウス分布で近似されない場合における相互情報量の現実的に有効な推定方法の検討，音源が複数ある場合への拡張，および環境中に設置した多種類のセンサを含む場合への適用が挙げられる．

文　献

[1] S. Becker and G. E. Hinton. A self-organizing neural network that discovers surfaces in r andom-dot stereograms. *Nature*, 355(9):161–163, 1992.

[2] S. Becker. Mutual information maximization: Models of cortical self-organization. *Network: Computation in Neural Systems*, 7(1), 1996.

[3] M. H. Coen. Multimodal integration - a biological view. *Proc. 17th Int. Joint Conf. on Artificial Intelligence (IJCAI'01)*, Vol. 2, pp. 1417–1424, 2001.

[4] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley-Interscience, 1991.

[5] J. W. Fisher III and Darrell. Speaker association with signal-level audiovisual fusion. *IEEE Trans. Multimedia*, 6(3):406–413, 2004.

[6] J. W. Fisher III, T. Darrell, W. T. Freeman, and P. Viola. Learning joint statistical models for audio-visual fusion and segregation. *Proc. Neural Information Processing Systems (NIPS'00)*, 2000.

[7] K. W. Grant and P. Seitz. The use of visible speech cues for improving auditory detection of spoken sentences. *J. Acoust. Soc. Am.*, 108(3):1197–1208, Sep 2000.

[8] J. Hershey, H. Ishiguro, and J. R. Movellan. Audio vision: Using audio-visual synchrony to locate sounds. *Proc. Neural Information Processing Systems (NIPS'99)*, 1999.

[9] E. Kidron, Y. Schechner, and M. Elad. Pixels that sound. *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR 2005)*, June 2005.

[10] D. Li, N. Dimitrova, M. Li, and I. K. Sethi. Multimedia content processing through cross-modal association˙ *Proc. ACM Int. Conf. on Multimedia 2003*, pp. 604–611, 2003.

[11] H. McGurk and J. W. MacDonald. Hearing lips and seeing voices. *Nature*, 264:746–748, 1976.

[12] H. J. Nock, G. Iyengar, and C. Neti. Assessing face and speech consistency for monologue detection in video. *Proc. ACM Int. Conf. on Multimedia 2002*, pp. 303–306, 2002.

[13] H. J. Nock, G. Iyengar, and C. Neti. Speaker localisation using audio-visual synchrony : An empirical study. *Proc. Int. Conf. on Image and Video Retrieval (CIVR2003)*, 2003.

[14] S. Shimojo and L. Shams. Sensory modalities are not separate modalities:p lasticity and interactions. *Current Opinion in Neurobiology*, 11(4):505–509, Aug 2001.

[15] M. Slaney and M. Covell. Facesync: A linear operator for measuring synchronization of video facial images and audio tracks. *Proc. Neural Information Processing Systems (NIPS'00)*, pp. 814–820, 2000.

[16] 山崎, 石川. センサフュージョン－実世界の能動的理解と知的再構成－. コロナ社, 1992.

[17] 石川 (編). 特集「センサフュージョン」. 日本ロボット学会誌, 12(5), 1994.

(平成 xx 年 xx 月 xx 日受付)

池田　徹志　（正員）

平 9 京都大学大学院工学研究科修士課程了．同年，三菱電機 (株) 入社．画像認識の研究開発に従事．平 15 大阪大学大学院工学研究科博士後期課程研究指導認定退学．平 17 大阪大学大学院工学研究科知能・機能創成工学専攻特任助手となり現在に至る．知覚情報基盤の研究に興味を持つ．

石黒　　浩　（正員）

平 3 大阪大学大学院基礎工学研究科物理系専攻修了．工学博士．同年山梨大学工学部情報工学科助手．平 4 大阪大学基礎工学部システム工学科助手．平 6 京都大学大学院工学研究科情報工学専攻助教授，平 10 同大学大学院情報学研究科社会情報学専攻助教授．この間，平 10 より 1 年間，カリフォルニア大学サンディエゴ校客員研究員．平 12 和歌山大学システム工学部情報通信システム学科助教授．平 13 より同大学教授．平 11 より，ATR 知能映像研究所客員研究員．現在，大阪大学大学院工学研究科知能・機能創成工学専攻教授および ATR 知能ロボティクス研究所客員室長．知能ロボット，アンドロイドロボット，知覚情報基盤の研究に興味を持つ．

浅田　　稔　（正員）

昭 52 大阪大学基礎工学部制御工学科卒業，昭 57 同大学院基礎工学研究科修了工学博士（大阪大学）．平 7 大阪大学工学部教授．平 9 大阪大学大学院工学研究科知能・機能創成工学専攻教授となり現在に至る．この間，昭 61 から 1 年間米国メリーランド大学客員研究員．知能ロボットの研究に従事．ロボカップ国際委員会プレジデント．平 17 年 9 月より、科学技術振興機構　戦略的創造研究推進事業（ERATO）「浅田共創知能プロジェクト」研究総括．

Abstract Previous approaches of sensor integration combined different kinds of sensors after feature extraction and abstraction. The problem in these approaches is correlated multimodal information between sensory signals is lost during independent abstraction process. This paper proposes a new approach that combines sensory signals from different kinds of sensors in the earlier stage. By formalizing sensor integration process as an optimization that maximizes mutual information between sensory signals, a moving signal source in a changing scene is detected by a heuristic search algorithm. As an example, experimental results of sound source detection with one video camera and one microphone are shown.

Key words Mutual information, perceptual information infrastructure, sensor fusion, sensor integration, sound source localization