相互情報量最大化に基づく異種センサの信号の適応的統合

○池田徹志 (阪大院)　石黒浩 (和歌山大)　浅田稔 (阪大院)

Adaptive Fusion of Sensor Signals based on Mutual Information Maximization

*Tetsushi IKEDA (Osaka Univ.), Hiroshi ISHIGURO (Wakayama Univ.) and Minoru ASADA (Osaka Univ.)

Abstract— Mutual information maximization is promising criteria for fusing signals of different type of sensors. In previous approaches, fusing algorithm is rigid and it is hard to apply, for example, to moving objects. In this paper, we propose adapting the way of fusing accoding to the nature of input signals. An example of adaptive fusion is provided.

Key Words: Mutual Information, Correlation, Adaptive Sensing, Audio, Visual, Sensor Fusion

1. はじめに

多様な環境をノイズ等に影響されず頑健に認識する方法の１つは，複数のセンサを用いて観測した結果を統合することである．従来のセンサ統合の研究では，各センサから得られる信号に対して個別に特徴抽出を行い，特徴の組み合わせに対してモデルを適用する手法が主流である．

環境を共有する複数のセンサから得られた信号は，同一の環境が異なる経路で観測されたものであると考えると，信号は特徴抽出以前の段階で既に互いに強い関係があることが期待できる．近年，信号のこのような関係に注目して統計的手法により統合を行う研究が行われている [1)2)3)]．

Becker[1)] は，２つのニューラルネット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，ニューラルネットの入力間の平行移動量等の興味深い特徴を抽出できることを示した．Hershey et al.[3)] は，音声と画像の強度の時系列の相互情報量を求めることにより，話者を画像上で特定できることを示した．また Fisher et al.[2)] は相互情報量最大化の規範で音声と画像に対する変換を学習し，話者の画像上での特定および指定された画像領域の情報を元に音声のフィルタリングを行った．

これらの研究の本質的な問題点は，信号の統合処理が固定的であることと考えられる．そのため例えば対象が画像上で移動する場合に対応するのは困難である．また入力信号によらず統合は常に画像上で全探索を行っており柔軟性に欠ける．

ノイズ等の影響を抑え，柔軟に統合を行うためには，統合の対象となる信号の性質に応じて統合の仕方を適応させるアプローチが有効と考えられる．本研究ではその例として，複数センサの信号間の相互情報量を求める際の時間窓長を適応的に変化させる手法を提案する．また例を通じて，音を発しながら移動する物体の画像上での同定を効果的に行うことができることを示す．

2. 音と画像の統合手法

音のパワーと画像の各画素の輝度値の時系列の相互情報量を求め，高い相互情報量を示した領域を音と関係の強い部分として抽出する．

2·1 音と画像の信号間の相互情報量の導出

時刻 t での音特徴を A(t)，時刻 t，位置 (x,y) での画像特徴を V(t)，H() をエントロピ関数とするとき，音と画像の相互情報量 I は以下で与えられる [3)]．

$$I(A(t);V(x,y,t)) = H(A(t)) + H(V(x,y,t)) - H(A(t),V(x,y,t)) \quad (1)$$

ここで，$p(A(t))$, $p(V(x,y,t))$, $p(A(t),V(x,y,t))$ が Gauss 分布に従うと仮定し，音特徴と画像特徴が共に１次元の量とすると，上記の相互情報量は以下のように表すことができる．

$$\frac{1}{2}\log\frac{\sigma_a\sigma_v}{\sigma_{av}} \quad (2)$$

ここで，σ_a，σ_v はそれぞれ音特徴および画像特徴の分布の分散，σ_{av} は両者の共分散である．一定長の時間窓 (長さ T) に関してこれらの分散と共分散を求めることにより，式 (2) を用いて相互情報量を求めることができる．

2·2 相互情報量を計算する時間窓長の適応

対象が画像上で移動しない場合には，統合を行う時区間の長さ T をできるだけ長くとることによりノイズ成分が除去され，音と画像の関係を精度良く抽出できる．しかし対象が移動する場合には，T が大きすぎると移動に追従できず，小さすぎるとノイズの影響を受けやすく，両者の関係を抽出することが難しくなり，トレードオフが生じる．

ここで各時点での相互情報量に基づいて T を適応させる．式 (2) に基づいて計算された相互情報量を I，音と画像の関係を十分に抽出できたと考えられる相互情報量の大きさを $\hat{I}$ とした時，次式にしたがって T を修正する．

$$\Delta T = \alpha(\hat{I} - I) \quad (3)$$

ここで α は定数である．

3. 実験

人が足音をたてて歩いている様子をカメラとマイク各１つで観測し，足音に基づいて画像上での人の位置を同定する例を取り上げる．画像の例を Fig.1 に，音のパワーの系列を Fig.2 に示す．画像の左端には PC のディスプレイのちらつきが写っており，足音とは無関係の動きをしている．時間窓長を固定した場合と，適応的に変化させた場合のそれぞれついて実験を行った．

画像の解像度は 160x120，サンプリング周期は 1/30 [sec] である．音はサンプリング周波数 16kHz で録音し，画像の周期に合わせて 1/30[sec] ごとに平均のパワーを求めたものを用いた．またこれらの実験は PC (PentiumIII 1GHz) 上で実時間で行うことができる．

(a) frame 200 (b) frame 380

Fig.1 Example images

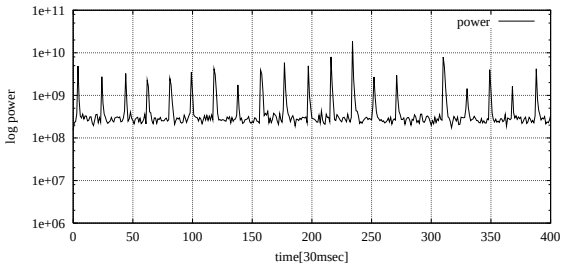
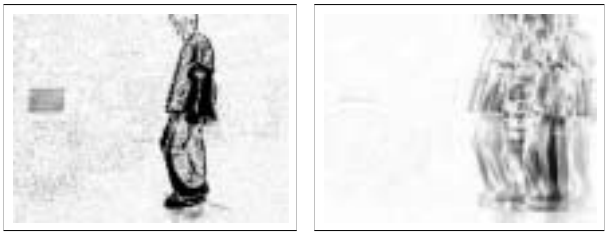


Fig.2 Audio power sequence

3·1 時間窓長を固定にした場合

時間窓長を固定にした際の結果を Fig.3 に示す．ここで濃い領域ほど相互情報量が大きいことを示している．Fig.3(a) は $T = 8$ とした場合である．人の位置は明確に検出されているが，画像左のディスプレイの部分や背景領域にノイズが現れている．Fig.3(b) は $T = 256$ とした場合である．ノイズは抑えられているものの，人の位置はぼけており正確に同定できていない．



(a) frame length = 8 (b) frame length = 256

Fig.3 Intensity of mutual information (fixed window length)

3·2 時間窓長を適応的に変化させた場合

各画素ごとに時間窓長を適応的に変化させた際の結果を Fig.4 に示す．$T = 16$ を初期値とし，$\hat{I} = 0.0005$, $\alpha = 4000$ とした．また時間窓長の最大値は 256，最小値は 8 とした．人の位置は明確に検出されており，また背景のノイズも抑えられていることが分かる．この時点での時間窓長を Fig.5 に示す．濃い部分ほど時間窓長が長いことを示す．

Fig.4 Intensity of mutual information (adaptive window length)

Fig.5 Window length

4. まとめ

本報告では，相互情報量の最大化の規範に基づき異種センサの信号間の統合を行う時に，統合方法を信号の性質に基づいて適応させる手法を提案した．例として相互情報量を計算する際の時間窓長を適応させることにより，対象が移動している場合でも，精度良く画像領域中の音源を抽出できることを示した．今後の課題は，統合する画像特徴や音特徴も適応的に選択してゆく，より柔軟なセンサ統合方式の構築である．

参考文献

1) S. Becker. Mutual information maxmization: Models of cortical self-organization. *Network: Computation in Neural Systems*, 7(1), 1996.
2) J. W. Fisher III, T. Darrell, W. T. Freeman, and P. Viola. Learning joint statistical models for audio-visual fusion and segregation. In *Advances in Neural Information Processing Systems*, 2000.
3) J. Hershey, J. Movellan, and H. Ishiguro. Looking for sounds: Using audio-visual correlation to locate sound sources. In *Proc. of Neural Information Processing Systems (NIPS'99)*. 1999.