人間行動観測に向けた複数種センサの定常性に基づくデータ処理

筑波大学大学院 システム情報工学研究科 ○服部 傑，亀田 能成，大田友一

Data Processing Based on Normality of Multimodal Sensors toward Human Action Recognition

○ Takashi HATTORI, Yoshinari KAMEDA, Yuichi OHTA
Graduate School of System and Information Engineering, University of Tsukuba

Abstract : This paper describes a method to extract significant data segments from continuous video and audio segments that include noise of daily life. It also discusses criteria to define resolution of feature vector units that can be easily utilized for human action recognition in sensor fusion approach.

1 はじめに

人間が住む環境自体を付加価値の高いものにし、より住みやすい社会にしていこうという動きが社会的に認められてきている。このような高付加価値型居住環境システムを完全に実現するには、居住空間すべてにおいて人間の行動観測を安定して行うことが必要になる。

人間の行動は運動を伴うこともあれば、音の発生を伴うこともある。また、観測によって得られる情報量は多いほど行動の認識や分類に役立つと考えられる。そこで我々は、受動的センサであるカメラやマイクロフォンを空間内に多数配置し、人間の行動観測を行うことを考えている。

本アプローチにおいて、まず問題になるのは、受動的センサから得られる情報からの有用な情報の抽出である。カメラやマイクロフォンは受動的センサとして多くの情報を獲得可能であるが、日常生活環境ではセンサ観測量にノイズが混在しやすい。これは環境中に定常的なノイズ発生源やゆるやかな変動要因が存在するためである。そこで我々は、ノイズの定常性に注目し、そのような環境でも有用な観測データを抽出する方法について本稿で述べる。

また、もう一つの問題は、センサフュージョンの方法である。我々は特定行動の認識ではなく、人間行動全体の観測を考えているので、その行動が映像的に観測できるか、ないしは音響的に観測できるのかを予見することはできない。そこで、本稿では、映像特徴量と音響特徴量を統一的に扱うための情報の粒度の均質化について述べる。

2 有意部分の抽出とデータ粒度の均質化

カメラやマイクロフォンのような、大量のデータを出力し続けるセンサでは、そのすべての出力について処理・記録を行うことは現実的ではない。そこで、意味があると見なせるデータ系列のみを抽出する必要がある。また、その一方で、センサからみた日常環境は静止した固定的なものには限らず、定常的なノイズや緩やかな変動要因を含んでいる。そこで本手法では、映像・音声データ各々に対し過去一定期間のデータを保持し、動的背景モデルを更新するとともに、そこから逸脱した部分を有意部分として抽出する。

2.1 M 推定を用いた背景モデルの推定と前景画像抽出

我々の想定している日常環境でのセンシングでは、室内であっても屋外の日照状況の変化や照明状況の変化を受けたり、家具や物品等の配置が時々変化したりするので、何らかの方法で背景画像を更新していく必要がある。そこで、ロバスト統計に基づく M 推定 [1] を用いて動的背景画像モデルを構築し、同時に前景の抽出を行う。この手法では、照明状況の変化・対象環境の変化に対して適応的に背景推定モデルを更新していくことができる。

2.2 定常ノイズモデルを用いた前景音像抽出

映像データに対する処理と同様に、定常音に対応した背景音モデルを定義し、前景音像を取り出す方法を用いる。具体的には、一定時間間隔で区切られた音波形データに対してフーリエ変換を行い、各周波数成分ごとに過去一定時間のデータによる平均と分散で背景音モデルを定義する。新たに入力されるデータのある周波数成分 λ に対し、式 1 で定義される式によって背景音声からの相違度 d_λ を計算する。

$$d_\lambda = (p_\lambda - \bar{m}_\lambda)^2 \times \bar{v}_\lambda^{-1} \quad (1)$$

ここで $\bar{m_\lambda}$ と $\bar{v_\lambda}$ はそれぞれ背景音モデルの平均と分散を表し、p_λ は入力音波形の λ 周波数成分の強さを

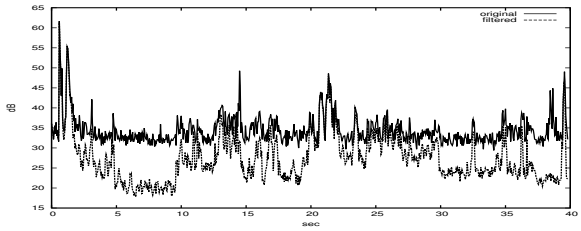

Fig. 1: 音声データ

示す。

この相違度 d_λ が閾値以上であったものを前景周波数成分として採用し、それ以外のデータを取り除く。その後逆フーリエ変換を行い、その強さによって前景音声かどうか判別を行う

本手法を適用した例を図 1 に示す。図中 original と表記した入力値は入力音波形を表す。本入力波形データでは、エアコンのファンや PC の排気ファンなどの音がする中で、発話が 9.5 秒〜30 秒付近でなされている。本手法を適用した結果が filtered である。発話区間の音圧が保存され、S/N 比が向上していることがわかる。

2.3 センサ特徴量の粒度の均質化

異種のセンサを統一的に利用する場合、各センサの出力する特徴量の時間方向の粒度と次元数の違いが問題となる。時間方向については、可聴域の音の観測は数十 KHz 程度で量子化されるのに対し、一般に映像は数十 Hz で量子化される。一方、人間行動観測が目的であるので、観測に必要な人間行動の時間的分解能についても考慮すべきである。我々は、センサと人間とのおおよその距離も考慮して、映像・音声の特徴量とも時間方向の粒度を 100msec 刻みに調整することにした。

また、各時間データセグメントにおけるそれぞれのセンサ観測量から得る特徴量次元数についても、その数が著しく異なるとセンサフュージョンが難しくなると予測されるので、其々の次元数をおおよそ同じオーダーに収めるものとする。

3 システム構成と実験環境

システムは、ネットワーク接続されたセンサ群と処理用 PC 群から構成されている。構成要素は以下の通りである。実験環境は、本学実験室とその周辺廊下を対象としている。各センサは図 2 のように配置されている。

- **ネットワークカメラ群**: 実験室天井と周辺廊下天井に計 40 台程度を設置する。これは、実験環境床面を死角無く撮影できる数である。
- **マイクロフォン群**: マイクロフォン群は実験室天井に設置する。これらは AD コンバータを用いて一台の処理用 PC に接続されている。
- **画像処理用 PC 群**: ネットワークカメラ群に対して画像取得・処理を行う。

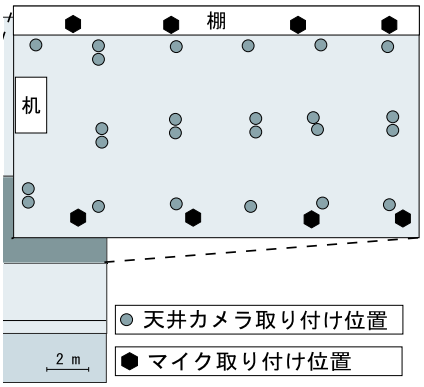

Fig. 2: センサ配置図

この環境中で、一人の人物が日常動作を行い、本手法による前景抽出を映像・音声について行った。

約 5 分間の間に、実験室脇の部屋から実験室への移動・実験室内の徘徊・戸棚の開閉・実験室からもとの部屋への移動をした。

映像は、各カメラの撮影領域を 25 分割し、RGB の 3 成分 75 次元を特徴量とした。また、音声は 10〜13 KHz の範囲から、27 個の代表周波数定義し、27 次元の特徴量とした。特徴量の時間方向の粒度はすべて 100 msec に統一した。その上で。有意な情報を含む区間の抽出を 1 sec 刻みのセグメント単位で行った。

その結果、映像は全 10185 セグメント中 1764 個が、音声は全 1686 セグメント中 826 個が有意なデータ区間と判別された。これは主観的に確認した結果とほぼ一致した。

4 まとめ

本稿では、環境中の定常性を考慮して、カメラ・マイクロフォンで観測した情報から、有意な部分を自動的に抽出する方法を述べた。また、その上で人間行動認識のためにセンサ間の粒度をそろえて処理する方法についても提案した。カメラ・マイクそれぞれの出力について動的背景モデルを構築・更新することによって、前景と見なせるデータのみを抽出可能である。今後は、得られた多次元特徴量セグメント群に対し、DP マッチングによって距離を定義し、その上で人間行動のクラスタリングや判別を行っていく予定である。

参考文献

1) 島井, 他: “ロバスト統計に基づいた適応的な背景推定法”, 信学論 D-II, Vol.J86-D-II, No.6, 2003