社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

人間行動観測のための視聴覚センサを用いた非定常部分の抽出

服部　傑†　亀田　能成†　大田　友一†

† 筑波大学大学院システム情報工学研究科　〒 305–8573 つくば市天王台 1-1-1
E-mail: †{hattori,kameda,ohta}@image.esys.tsukuba.ac.jp

あらまし　高付加価値型住居環境システムを実現するためには、人間の行動観測を安定して行う必要がある。我々はその一環として、多様なセンサを用いた人間の行動認識の実現に向けて研究を行っている。本稿では、環境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、有意な部分情報を自動的に抽出する手法について述べる。また、その上で人間行動認識のために各センサのデータ出力の粒度をそろえる方法についても提案する。本手法ではカメラ・マイクロフォン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する。実際に実験環境において日常行動の観測を行い、本手法を適用した。実験結果をふまえて、行動認識処理に必要な時系列情報間の類似度の定義について考察する。
キーワード　センサフュージョン，定常ノイズ，行動認識，動的背景モデル

Abnormality Extraction by Utilizing Video and Audio Sensors toward Human Action Recognition

Takashi HATTORI†, Yoshinari KAMEDA†, and Yuichi OHTA†

† Graduate School of Systems and Information Engineering, University of Tsukuba　1-1-1 Tenoudai, Tsukuba, 305-8573 Japan
E-mail: †{hattori,kameda,ohta}@image.esys.tsukuba.ac.jp

Abstract　A stable and robust human action recognition is necessary for realization of high-value added living environments. Our purpose is to develop a sensing system for human action recognition by utilizing a number of multimodal sensors. In this paper, we describe a method to extract significant data segments from continuous video and audio data that include noise of daily lives. This method extracts the significant data segments by adaptive background model that can deal with stationary and ambient noise. We also discuss criteria to determine resolution of feature vector units so that they can be easily utilized for human action recognition in sensor fusion approach. We apply this method to some basic actions in an experimental environment. Finally, we also discuss adequate definition of similarity between the time-series data that are extracted by our method.
Key words　Sensor fusion，stationary noise, action recognition, adaptive background model

1. はじめに

インテリジェントルームやアウェアホーム、インテリジェント在宅福祉、ビデオサーベイランスなど、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に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部分情報の抽出である。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入しやすい。これは環境中の背景とみなされる構成要素の中に、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、背景の動的特性に注目し、そのような環境でも

非定常部分である有用な部分情報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。ある特定の目的に対しては、認識対象に関する知識に基づいて各センサの出力を処理し、統合するアプローチが効果的であるということが分かっている (西口ら [2]、大西ら [3])。我々は特定行動の認識ではなく、人間行動全体の観測を目標として考えているので、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、観測された時系列情報を判別するための距離の定義について考察する。

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記録をしたり高次処理をする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。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とは限らず、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器具の変化の影響を受けたり、家具や物品等の移動にも影響を受けたりするので、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。

以下、M 推定を用いた背景画像推定について簡単に説明する。M 推定はロバスト推定法の一つであり、累積誤差の最小化問題 (式 (1)) での誤差基準を式 (2) のようなロジスティック関数で定義し、外れ値の影響を一定以下に抑える。ここで、θ_t はある時刻 t での背景推定値を、x_t は同時刻に得られる画像の入力値をあらわし、ϵ_t はその際の誤差をあらわしている。

$$min \sum_t \rho(\epsilon_t) = min \sum_t \rho(x_t - \theta_t) \tag{1}$$

$$\rho(x) = log(cosh(\frac{x}{a})) \tag{2}$$

M 推定を用いた適応的な背景画像推定では，背景の時間的変化に応じて最急降下法によってそのモデル θ を更新する (式 (3))．ここで、σ は累積誤差に対する適応率を、E_t は時刻 t までの累積誤差をあらわす。

$$\theta_t = \theta_{t-1} + \sigma \frac{\partial E_t}{\partial \theta} \tag{3}$$

ここで、古いデータの影響が残り続けることは、適応的な背景更新にとっては好ましくないと考えられる。そのために，古いデータの重みを指数関数的に減衰させながら累積誤差を逐次的に更新していく（式 (4)）．この式で，$\alpha(0 \leqq \alpha \leqq 1)$ は記憶率を表し，E_t は t 枚の画像を用いた重みつき累積誤差を意味する．

$$\begin{aligned} E_t &= \sum_{l=0}^{t-1} \alpha^l \rho(\epsilon_{t-l}) \\ &= \rho(\epsilon_t) + \alpha E_{t-1} \end{aligned} \tag{4}$$

式 (4) を式 (3) に適用すると、背景モデルの推定値 θ_t は式 (5) のように求まり、これによって背景モデルを逐次的に更新していく。

$$\begin{aligned} \theta_t &= \theta_{t-1} + \sigma \frac{\partial E_t}{\partial \theta} \\ &= \theta_{t-1} + \sigma \left[\frac{\partial \rho(\epsilon_t)}{\partial \theta} + \alpha \frac{\partial E_{t-1}}{\partial \theta} \right] \end{aligned} \tag{5}$$

なお、本稿における適応率 σ および記憶率 α は、実験環境で大きな日照変動をもたらす日没前後に撮影された画像における背景画素値の変化から、実験的に決定した。背景モデルをこのように更新した後、入力画素値 x_t と θ_t とを比較し、その差が閾値以上となる領域を前景とする。

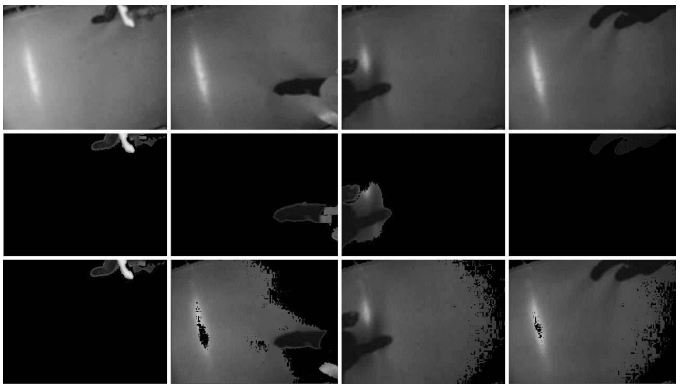

図 1　前景抽出結果
上段：入力画像
中段：本手法による前景抽出結果
下段：背景更新を行わない場合の前景抽出結果

図 1 は、あるカメラにおいて動的に背景モデルを更新しながら前景抽出を行った結果である。上段が入力画像で、中段が本手法により前景部分を抽出した画像である。比較のため、下段に背景更新を行わなかった場合の前景抽出結果を示す。約 3 時間にわたるデータに対して、背景更新と前景抽出をした。左から順に約 1 時間おきの画像となっていて、時刻の移り変わりとともに全体的な明るさが変化しているが、背景更新によって正しく前景だけを抜き出せていることがわかる。

2.2 定常ノイズモデルを用いた前景音像抽出

音像についても、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、前景を取り出す方法を用いる。背景音スペクトルを利用し、前景音抽出を行うという手法は、Boll ら [4] によって提案され、その後様々な改良がなされている [5]。本手法では、この SpactralSubtraction 法に M 推定を用いて背景音スペクトルを推定し、それを用いた前景音抽

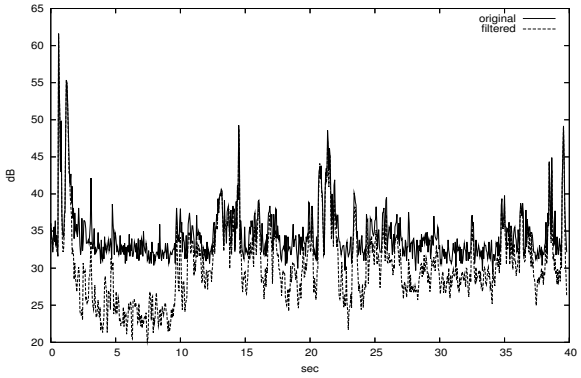

図 2　音声データ

出を行う。

具体的には、下記の手順で処理を行う。あるマイクロフォンから得られる音波形データを一定時間間隔で区切り、フーリエ変換を行う。そして各周波数成分ごとに、映像の場合と同様に M 推定によって背景音の強度を推定する。この推定背景周波数成分強度と入力値の一定時間の分散によって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 (6) によって背景音声からの相違度 d_λ を計算する。

$$d_\lambda = (p_\lambda - \theta_\lambda)^2 \times \bar{v_\lambda}^{-1} \tag{6}$$

ここで θ_λ と $\bar{v_\lambda}$ はそれぞれ背景音モデルの M 推定による推定値と該当周波数 λ における入力値の分散を表し、p_λ は入力音波形の λ 周波数成分の強さを示す。

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。これが、音像での前景部分となる。その後逆フーリエ変換を行い、フィルタリングされた時系列の音波形データとする。

本手法を適用した例を図 2 に示す。図中 original と表記した入力値は入力音波形を表す。横軸は時間を表し、縦軸は入力音波形の音圧を表している。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、そのほかの区間と発話区間との S/N 比が 15.2 dB から 22.3 dB に向上した。

2.3　センサ特徴量の粒度の均質化と有意データ節の抽出

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は、しばしばそれらを用いた認識器の性能に影響を与える。一般的なセンサの出力には、時間方向の分解能と、一回のセンシングで得られる観測値自体の分解能がある。具体的には、カメラであれば、前者はフレームレート，後者は解像度および各色成分の分解能であり、マイクロフォンであれば、前者はサンプリング周波数，後者は量子化ビット数である。

一般に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、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、映像・音声の特徴量とも時間方向の粒度を 100msec 刻みに調整している。

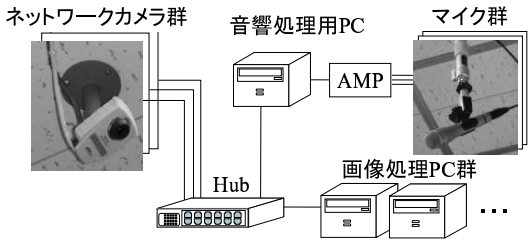

図 3　システム構成

一方、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は、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想されるので、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。

具体的には、映像は各カメラの撮影領域を 25 分割し、RGB の 3 成分の計 75 次元を特徴量とした。また、音声は 10〜13 KHz の範囲から、27 個の代表周波数を定義し、27 次元の特徴量とした。

我々の研究では、一定の長さを持つ時系列データの節を単位として、その後の判別処理に用いることを考えている。そこで、上述の手法で取得した前景時系列データに対して、有意データ節の抽出をする。具体的には、入力となる時系列データから前景時系列データを取り出し、定常状態から外れていると判断される区間を含む一定時間長を有意データ節として抽出する。

3.　システム構成と実験環境

3.1　システム構成

システムは、図 3 のように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 4・5・6 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 35 台を設置する。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**: マイクロフォンは 8 台を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・処理をする。

今回の実験では、フィルタリングを含む処理はすべてオフラインで行っている。

3.2　入力データ

本システムの設置環境、特に実験室では、日常的に以下のような作業（イベント）が発生する。

- 棚に入っている物品の持ち出し・返却
- 物品の探索
- 印刷・印刷物の回収
- 撮影実験・作業

環境中では、これらの作業の発生頻度は日に数回程度である。

図 4　センサ群配置:実験室

図 5　センサ群配置:廊下

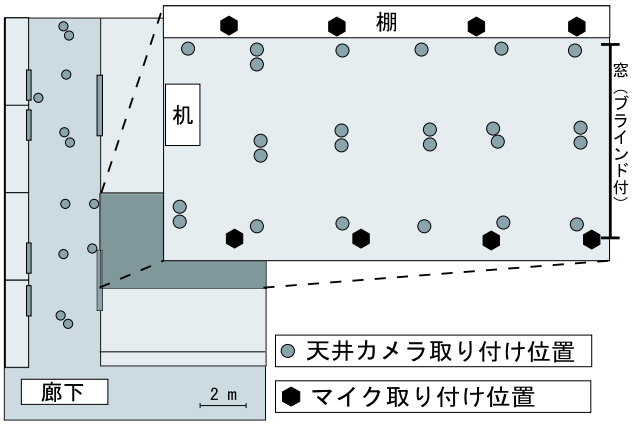

図 6　センサ配置平面図

本来は設置環境から長期的にデータの収集を行い、それらに対して処理を適用すべきであるが、現在の実験システムの制約から、今回は実験のために意図的に数分おきに上記のような作業を反復して行い、それに対して観測を行った。

具体的には、約 3 時間にわたって下記の作業を 5 回ずつ反復して行った。

- 機材棚からのソフトウェアの持ち出し、および返却
- 機材棚からの機材の持ち出し、および返却
- 機材棚の探索（物色）
- 印刷・印刷物の回収

それぞれの試行において、自然な日常行動に近い入力データを得るために、行動経路や振る舞いなどは統一しなかった。作業を行う人物は一人であったが、上着を二種類用意し適宜取り替えた。また、実験環境に対して特に出入りの禁止をしなかったので、意図的に行ったもの以外の日常行動も入力として観測されている。

表 1　抽出された有意データ節数（カメラ）

CAM	有意節数	CAM	有意節数	CAM	有意節数
1	56	13	834	25	1198
2	231	14	405	26	897
3	891	15	352	27	983
4	631	16	246	28	686
5	471	17	283	29	863
6	361	18	161	30	90
7	333	19	585	31	834
8	1749	20	20	32	715
9	797	21	9	33	171
10	4861	22	494	34	606
11	1598	23	1093	35	192
12	934	24	348		

3.3　有意データ節 (セグメント) 抽出結果

次に、取得したデータに対して、有意データ節の抽出を行った。節の抽出の為のデータ処理は各センサ単位で行った。今回は、有意データ節の時間長を 1 sec 単位とし、その中に定常状態から外れていると判断されるデータが含まれる区間を有意データ節として出力した。

カメラからの入力に関しては、各フレームにおいて前景と判断された画素の数が全体の 1 %を超えるものを有意画像と定義し、それを含む区間を有意データ節とした。

また、マイクロフォンからの入力は、背景音像による前景音像抽出フィルタを適用した音波形データを処理対象とした。この音波形データに閾値以上の強さの波形が観測されるものを含む区間を有意データ節と定義して抽出した。本実験では、有意データ節か否かの判定をする閾値として、各マイクロフォンでフィルタリング後の音波形の最頻値付近の値を用いた。

抽出された有意データ節数を表 1 および表 2 にまとめる。本実験では 2 時間 48 分 13 秒分のデータを取得しており、抽出される候補となる区間は 10093 個あった。CAM1〜12 までが廊下に設置されており、CAM13 以降のカメラが実験室天井に設置されている。

CAM10 の出力有意データ節数が際だって多い。原因として、このカメラが本研究室入り口を撮影しており、入り口に置かれていた靴が長時間にわたって前景と判定され続けたことが考えられる。本実験では、日照変化にあわせて背景推定用の学習パラメータを調整していたが、文献 [1] で取り上げているように学習パラメータを適応的に調整する手法を取り入れることで、今後はこのような放置された物体に対処できると考えられる。

CAM20・21・30 の出力有意データ節数が他と比較して極端に少ないのは、これらのカメラが実験室の奥の方を撮影しており、実験中に当カメラの撮影領域に人間が立ち入ることが少なかった為である。CAM18・33・35 も同様に、大量の物品が置いてあり通行が難しい領域を視野に含んでいるため、相対的に出力有意データ節が少なくなった。

次に、マイクロフォンから出力された音波形データから抽出

表 2　抽出された有意データ節数（マイクロフォン）

MIC	有意節数	MIC	有意節数
1	1164	5	761
2	935	6	1949
3	1239	7	725
4	853	8	1327

した有意データ節の数について考察する。

対象実験環境では音に関連するノイズ源として、エアコンの稼動音・PC や各種機器の稼動音が存在する。このノイズ音は、普通に会話する程度の音と大差ない程度の大きさがある。逆に、作業や入退室に伴う扉・棚の開閉音はそれらと比べて倍以上の大きさの入力として観測されている。

各マイクロフォンでの出力有意データ節数はおおむね 1000 前後に収まっている。これは、棚・扉の開閉に関連する作業を観測可能なカメラ (CAM13:入り口扉付近, CAM22:機材棚, CAM31,32:ソフトウェア棚) の出力有意データ節数に近い値を示している。

MIC6 の出力有意データ節数が他と比べて大きな値となっている。しかし、入力音波形の傾向が他のマイクロフォンと著しく異なっているわけではないので、閾値の設定値が大きく影響していると考えられる。適応的に閾値を決定するなどの対処は、今後の課題としたい。

3.4　有意データ節間の類似尺度

抽出された各有意データ節は、多次元の時系列データである。これらをクラスタリングすることによってイベントの判別をするためには、時系列データ間の類似度を何らかの形で定義する必要がある。

単純に時系列データ同士をマッチングする場合、似たイベントで有意データ節内の時間がずれていと、データ間の距離が大きくなってしまう。言い換えれば、時系列データをマッチングする場合には、データ間の位相のずれが問題となる。

位相のずれを吸収するような適切な類似尺度を用意することで、それに基づいたクラスタリングとイベント判別が可能になる。しかし、選択した類似度の尺度によって判別結果が大きく左右されることが考えられる。

そこで本稿では、以下の三種類の類似尺度について検討を行った。ここで、ある時系列データ $\mathbf{a}$ は、N 次元・T 時間の大きさを持つ多次元データであり、次元 n・時刻 t におけるデータを a_{nt} とあらわす。また、時系列データ $\mathbf{a}$ と $\mathbf{b}$ の DP マッチングを $DP(\mathbf{a},\mathbf{b})$ と表記する。

- 多次元並列 DP マッチング（DP_1）：

多次元データの各次元で DP マッチングを行い、それらの平均値を多次元時系列データ間の類似度として定義する。具体的には、データ $\mathbf{a}$・$\mathbf{b}$ 間のある次元 n での類似度を、a_{nt_a} と b_{nt_b} におけるコストを式 (7) とする DP マッチング $DP(\mathbf{a_n},\mathbf{b_n})$ で定義し、データ $\mathbf{a}$・$\mathbf{b}$ 間の類似度を式 (8) で定義する。

$$cost(a_{nt_a}, b_{nt_b}) = |a_{nt_a} - b_{nt_b}| \tag{7}$$

$$d_{DP_1}(\mathbf{a},\mathbf{b}) = \frac{\mathbf{1}}{\mathbf{N}} \sum_{\mathbf{n=1}}^{\mathbf{N}} \mathbf{DP_1}(\mathbf{a_n},\mathbf{b_n}) \tag{8}$$

- 多次元 DP マッチング（DP_N）：

多次元空間のユークリッド距離をコストとする DP マッチングによって、多次元時系列データ間の類似度を定義する。具体的には、DP マッチングで a のある時刻 t_a と b のある時刻 t_b におけるコストを式 (9) のように定義し、それを用いた DP マッチングによって式 (10) のように類似度を定義する。

$$cost(\mathbf{a_{t_a}},\mathbf{b_{t_b}}) \;=\; \left(\sum_{n=1}^{N}(a_{it_a} - b_{it_b})^2\right)^{\frac{1}{2}} \tag{9}$$

$$d_{DP_N}(\mathbf{a},\mathbf{b}) \;=\; DP_N(\mathbf{a},\mathbf{b}) \tag{10}$$

- RunningSpectrum（RS）：

各次元の時系列データ $\mathbf{a_n}$ をフーリエ変換し、次元数 $N \times$ 周波数成分数 L の多次元データ間のユークリッド距離を多次元時系列データ間の距離として定義する。具体的には、データ $\mathbf{a}$ の RunningSpectrum $\mathbf{A_{RS}}$ を式 (11) および (12) によって求め、式 (13) によって類似度を定義する。ここで、λ は FFT によって得られたある周波数を表す。

$$\mathbf{A_{RS}} \;=\; [\mathbf{A_1},\mathbf{A_2},\cdots,\mathbf{A_N}] \tag{11}$$

$$\mathbf{A_n} \;=\; FFT(\mathbf{a_n}) \tag{12}$$

$$d_{RS}(a,b) = \left(\sum_{n=1}^{N}\sum_{\lambda=1}^{L}(A_{n\lambda} - B_{n\lambda})^2\right)^{\frac{1}{2}} \tag{13}$$

これらの傾向を分析するために、本実験で得られた有意データ節に対してそれぞれの手法によって類似度を求めた。図 7 は、CAM13 で得られた有意データ節群で得られた一つ目の有意データ節に対して、それ以外の有意データ節への類似度をプロットしたものである。横軸は有意データ節番号を表し、縦軸は有意データ節間類似度を表す。なお、このグラフは類似度の変化の大きい一部の有意データ節との比較結果を拡大したものであり、三種の手法によって得られた数値を比較しやすいようスケールを変えてある。それぞれ、DP_1 が多次元並列 DP マッチング、DP_N が多次元 DP マッチング、RS が RunningSpectrum に対応している。また、CAM13 から得られた全有意データ節間の類似度を計算し、二次元プロットしたものを図 8 から 10 に示す。

各グラフから、三手法ともほとんど出力する値の傾向に違いがないことが分かる。一部の有意データ節で、注目有意データ節からの類似度の順が入れ替わっているものが存在する。例えば、有意データ節 705 番前後では三手法でピークの位置が異なり、また有意データ節 750 番付近では二カ所あるピークの形状が若干違っていることが分かる。これらの違いは、DP_1 に対してその他の手法が各次元を独立に扱っていないことが原因と考えられる。

三種法で二つの有意データ節間の類似度を計算するのにかかる処理時間は、DP_1 が平均 800 msec 程度・DP_N が平均 650 msec 程度・RS が平均 220 msec 程度であった。このとき

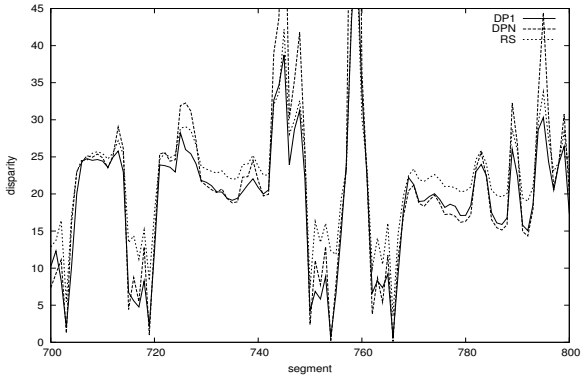

図 7　マッチング結果

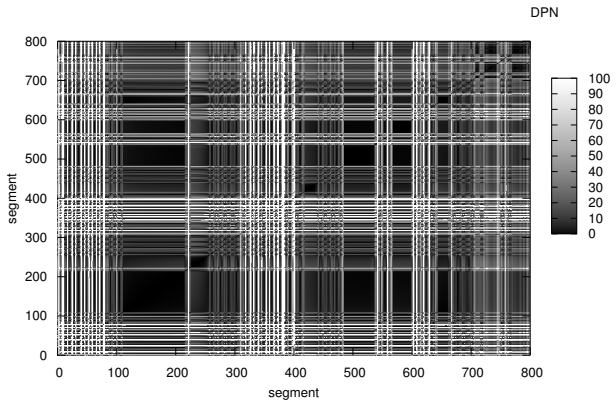

図 9　マッチング結果:DP_N

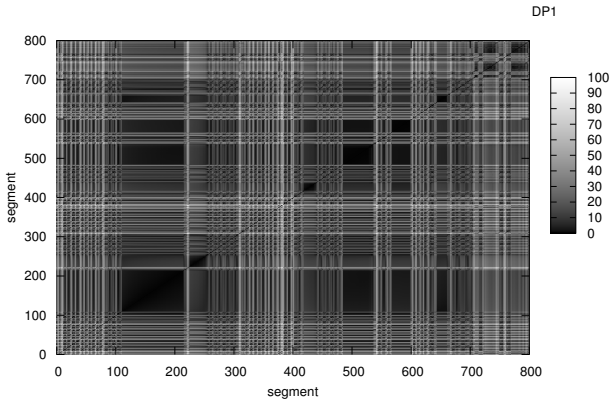

図 8　マッチング結果:DP_1

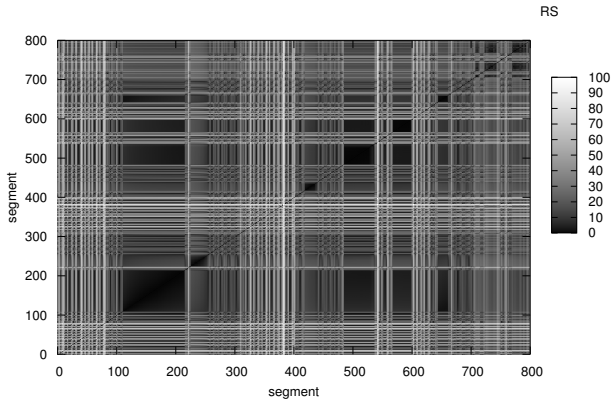

図 10　マッチング結果:RS

$N = 75$、$T = 10$(100 msec 刻みで 1 秒間) で、Pentium4 3.0 GHz の CPU を搭載する PC 上で処理を行った。

RunningSpectrum に基づく手法では類似度が多次元空間のユークリッド距離として定義される。つまり、時系列データ間の類似度が metric として定義されるために、その後の処理での扱いが比較的容易になると期待できる。さらに、処理時間が短くすむという利点もあるため、今後は RunningSpectrum に基づく類似度を利用する。

4. ま と め

本稿では、環境中におけるノイズ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手法について述べた。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・マイクロフォン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。さらに、実際に実験環境において日常行動の観測を行い、本手法を適用した。その上で、判別などの後処理を行うための時系列情報間の距離の定義について考察した。

文　　献

[1] 島井 博行，栗田 多喜夫，梅山 伸二，田中 勝，三島 健稔: “ロバスト統計に基づいた適応的な背景推定法”, 電子情報通信学会論文誌 D-II, Vol.J86-D-II, No.6, pp.796-806, 2003.

[2] 西口 敏司, 東 和秀, 亀田 能成, 角所 考, 美濃 導彦, “講義自動撮影における話者位置推定のための視聴覚情報の統合,” 電気学会論文誌C, vol.124, no.3, pp.729-739, 2004.

[3] 大西 正輝，影林 岳彦，福永 邦雄，“視聴覚情報の統合による会議映像の自動撮影，” 電子情報通信学会論文誌 (D-II)，vol.J85-D-II, no.3, pp.537-542, 2002.

[4] S.F.Boll: “Suppression of acoustic noise in Speech using Spectral Subtraction”, IEEE Trans. Acoust. Speech Signal Process., vol.27 No.2, pp.113-120, 1979.

[5] 岡崎 雅嗣，国本 利文，小林 隆夫: “信号の定常性を考慮したスペクトルサブトラクション法”, 電子情報通信学会技術報告 SP, 91-2003, pp.41-46, 2003.