連続DPによるスポッティングに基づく時系列画像認識

2000年2月提出

西村 拓一

もくじ

1 序言 **1**

1.1 はじめに 1

1.2 HMMとDP 2

1.3 連続DP 4

1.4 動画像を用いたジェスチャ認識 5

1.5 時系列画像を用いた移動ロボットの自己位置推定 6

1.6 時系列データからの類似時系列の検出 7

1.7 本論文の構成 8

2 動きと形状特徴を用いた連続DPによるジェスチャ認識 **9**

2.1 はじめに 9

2.2 動きと形状特徴 10

2.2.1 動きと形状特徴の概要 10

2.2.2 動き特徴 11

2.2.3 形状特徴 13

2.2.4 手領域の抽出 14

2.2.5 輪郭長による重み 14

2.2.6 手の位置による重み 16

2.3 連続DPによるスポッティング認識 17

2.4 動き特徴の評価 19

2.4.1 実験方法 19

2.4.2 実験結果 20

2.4.3 全方位視野を用いた認識 22

2.4.4 実験結果 23

2.4.5 実時間ジェスチャ認識システム 24

2.5 動きと形状特徴特徴の評価 26

2.5.1 実験方法 26

2.5.2 実験結果 26

2.6 まとめ 31

3 ジェスチャのオンライン教示 32

3.1 はじめに 32
3.2 モデルの切り出し 33
3.2.1 条件設定 33
3.2.2 モデル切り出し 33
3.2.3 パラメータ決定のための実験 34
3.2.4 実験結果 36
3.3 しきい値の決定 37
3.3.1 正規化距離の導入 37
3.3.2 パラメータαの決定 38
3.4 オンライン教示システムの評価 39
3.4.1 オンライン教示実験 39
3.4.2 実験結果 40
3.5 モデル平均化法 41
3.5.1 モデルの平均化手法の概要 42
3.5.2 連続DPによる対応づけ 42
3.5.3 モデルの平均化処理 43
3.5.4 平均化情報を利用した局所距離 43
3.6 モデル平均化法の評価 44
3.6.1 実験方法 44
3.6.2 実験結果 44
3.7 まとめ 45

4 非単調連続DP 47

4.1 はじめに 47
4.2 非単調連続DP 48
4.2.1 定式化 48
4.2.2 正規化係数αの時間可変化 52
4.3 戸惑い動作の認識 54
4.3.1 背景差分による低解像度特徴 54
4.3.2 戸惑い動作の認識実験 54
4.3.3 実験結果と考察 56
4.3.4 実時間ジェスチャ認識システム 58
4.4 移動ロボットの自己位置推定 58

4.4.1 非単調連続DPの必要性 58
4.4.2 従来の特徴量の問題点 59
4.4.3 回転不変な特徴量の概要と定式化 60
4.4.4 移動ロボットを用いた実験 61
4.4.5 実験結果と考察 64
4.5 まとめ 68

5 重み減衰型RIFCDP **69**
5.1 はじめに 69
5.2 重み減衰型RIFCDP 70
5.2.1 RIFCDPの問題点 70
5.2.2 重み減衰型RIFCDPの概念 71
5.2.3 定式化 74
5.2.4 類似区間検出法 75
5.3 重み減衰型ウインドウの特性 76
5.4 ジェスチャ動画像の検索 80
5.4.1 実験方法 80
5.4.2 実験結果 81
5.5 まとめ 82

6 緒言 **85**

第 1 章

序言

1.1 はじめに

我々の持つ五感のうちで，人間が外部から受けとる情報の80%が視覚によるものといわれ，それゆえ, 知覚の中で最も大きな役割を演じているのは視覚であるといえる．実際，人間の生活環境は3次元空間であり，これを把握するには二次元的かつ非接触な視覚を用いる方が，一次元の聴覚や嗅覚および接触が必要な触覚や味覚より高速で多量の情報を収集できる．従って，人間と同様に視覚情報を取得し，これを認識する機械が実現できれば，人間の行っている情報処理の多くの部分を自動化できる．この目的を実現するために，自動的に画像を解析し，認識するコンピュータビジョンの研究が1950年代ごろから活発に行われている [1]-[11].

既に取り組まれているコンピュータビジョンの課題は，航空写真や，人体のX線写真，細胞の顕微鏡写真など人間が直接見る事ができない画像の解析から，人間の生活環境や屋外情景を撮影した画像の理解など多岐にわたる．しかし，これらいずれの課題についても，70年代以前にはある一つの時点で得られる画像(静止画) を様々な手法で解析することに重点がおかれていた．特に, 外界を理解するという課題では, 2 次元の画像情報からいかに 3 次元の世界を復元するかが問題であり, 物体表面の反射特性から形状を推定 (shape from shading) する研究や両眼視によって距離情報を獲得する研究が行われていた.

その後1980年ごろになって，カメラおよび計算機の能力向上に伴い，動画像を用いて対象の「動き」の理解を目指す研究が盛んになってきた．この動画像処理では，時間の変化とともに得られる連続した複数の画像系列，つまり時系列画像を扱う．時系列を扱うことで, 対象の動きをも推定する必要が生じ問題が複雑となる一方，対象の切り出しや3次元形状復元に有利となる場合もある．2次元の時系列画像から3次元の物体の動きを推定する場合，まず，画像から変化部位を検出し，画像間の対応づけを行う．このとき，対象上の複数個の点の対応づけが必要となるが，局所的な特徴だけでは，一意に決定できない事が多い．特に, 物体の動きに伴って見え隠れする場合の対応付けは困難である．そこで，取り扱う世界および対象の形状および運動のモデルを導入する．これによって, 特徴点の拘束条件を導入でき安定した動き推定が実現できる．例えば, 剛体が運動している場合は，同一剛体の上の各点は共通した動きとなり，各

点が融合したり分離しない．また，質量を持つため，速度や加速度の上限も存在する．これらの仮定により，物体の3次元運動を推定する研究が行われている[2].

しかし，従来の研究では,各時点での対象の位置，向き速度などを「計測する」ことが目的とされることが多かった．一方, 計測した対象の位置などの時間的な変化を見分けることが必要となる場合もある．例えば, 人のジェスチャを認識する場合では,各時点での人物の姿勢だけでなく，姿勢の時間的な変化を識別する必要がある．これは，長時間のジェスチャ動画像中から繰り返し出現する類似した動きを抽出し，対象人物の癖を把握するという課題の場合も同様である．このように，動画像から得られる時系列データから，あるパターンを認識すること（時系列パターン認識）を本論文の目的とする．

1.2 HMMとDP

一般に時系列パターン認識の処理は次の3つのステップからなる．

[1] 対象物体を的確に表す特徴量を取り出す処理

[2] 取り出した時系列の特徴量を基にして，各カテゴリのモデルを作る処理

[3] 新たに入力される時系列データとモデルとを比較する処理

この処理のステップ2では, 認識したいすべてのカテゴリについてのモデルを作成し，3番目の処理にて，最も適合するカテゴリを認識結果とする．ステップ2，3では，通常，時系列パターンの時間的,（特徴量の作る）空間的な変動を吸収，正規化することが要求される．なぜなら，人物の自然なジェスチャや動きなどでは，時間的なスピードの変動や，姿勢などの特徴量の変動が避けられないためである．従って，時間的空間的にある程度異なるパターンでも，同一カテゴリに分類する能力が必要となる．

この認識手法として多く用いられているのは，Dynamic Programming (DP:動的計画法 [12][15]) とHidden Markov Model(HMM [13][14]) である．DPでは，時系列データそのものをモデルとし，入力の時系列との距離を最小とするよう入力を伸縮させる．このとき，ありえる伸縮の組み合わせすべてを全数探索するのではなく，漸化的に局所的な距離を最小化することで計算量を低減している．伸縮の範囲は音声やジェスチャへ適用する場合，1/2〜2倍とすることが多い．

一方, HMMでは，いくつかの「状態」を考え，あるカテゴリに属するパターンは「状態の間を，事前に求めた遷移確率で遷移」することで表現できるとしている．この各状態では，出力確率が定義されており，これに応じて出力が選択される．入力データとモデルとの類似度は，入力と同じ出力となる状態遷移すべての中から，累積確率が最大となるものとする．

HMMのモデルを決定するためには，これらすべての確率を推定する必要があるため，多くの（実用上，数千個の）学習データが事前に必要となる．しかし，DPでは時系列データそのものをモデルとできるため，一つの学習データでモデルを作成できるという特長がある．もちろん，十分な量の適切な学習データが用意可能な状況では，HMMはDPより高い認識能力を示す場合がある．これは，モデルとしてHMMがDPを含んでいるためである．

以上の議論の他に，まだ重要なDPとHMMの相違点がある．それは,「入力フレーム毎のスポッティング認識」が実現可能であるかという点である．音声や動画像のような時系列データを一定時間（例えば、10 msecや30msecなど)毎に分析や特徴抽出を行なうとき、この一定時間毎の特徴ベクトルをフレーム特徴と呼ぶ。また、このフレーム特徴の入力ごとにシステム全体の処理が同期して進行するとき、この処理をフレーム毎の処理と呼ぶ。さらに，入力された一連の時系列データからマッチングに最適な部分区間を自動的に切り出して認識することをスポッティング認識と呼ぶ．実時間の認識システムでは，入力フレーム毎のスポッティング認識ができることが望ましい．例えば，ジェスチャのスポッティング認識が実現できると，動作者は動作の開始や終了を意識せずに様々な動作を連続的に行うことが許される．つまり，動作者は自然な動きを拘束されずに，モデルとして登録したパターンのみを認識させることができる．このフレーム毎のスポッティング認識を漸化的に低計算量で実現できるのがDPの1つの特徴である．

HMMでもオフラインでのスポッティング認識は可能である．ここで言うオフラインのスポッティング認識とは，一定区間の入力が終了したあとで，その区間のどこに注目するカテゴリのものが存在しているかを認識する機能である．しかし，入力フレーム毎に，すなわち，各時刻で切り出しと認識を同時的に行なうという 入力フレーム毎のスポッティング認識はHMMでいろいろ試みられているが，現在まで成功しているとはいいがたい．HMMでスポッティングを行う確実な方法は，入力時系列から区間長を変化させて複数個の部分区間を切り出し，それらすべてについてモデルとの距離を計算することであるが，莫大な計算量となり非現実的である．

以上をまとめると，DPはHMMと比較して以下の2点の特長を持つ．

[1] モデルの獲得のために必要なデータが少ない（1個でも可能）．

[2] 入力フレーム毎のスポッティング認識が可能．

先に示した時系列パターン認識の処理の2番目のステップ（モデルの作成）では，あらかじめ認識の対象であるパターンが分かっている場合とこれが指定されていない場合とで，認識手法に対する要求が大きく異なってくる．前者の場合は，例えば「さようなら　というジェスチャを認識しなさい」などの様に具体的に認識したいカテゴリが明示されている．この場合，大量のデータをあらかじめ取得できればHMMの適用も可能となろう．しかし，人物によって大きく異なると考えられるジェスチャの認識の場合では，数個のデータからモデルを作成できるDPの方が，逐次モデルを教示でき動作者に迅速に適応できるという利点がある．一方，後者の場合では,「この時系列データ中には，どのような類似した部分系列がどの時点で現れるか」などの様な課題が与えられる．この場合，あらかじめ認識カテゴリのデータを大量に用意することは不可能であり，HMMのモデル作成は困難となる．つまり，HMMでは切り出されていない時系列データ中から任意の長さの類似区間を検出することは難しいといえよう．これに対してDPの場合は，先に挙げた2点の特長により，この問題に対処できる．第1の特長により，時系列データをそのままモデルとすることが可能となり，さらに第2の特長により，モデルと入

力データとの間で，類似区間の始端と終端を決定することが可能となる．次節では，DPを時系列データ認識のために改良した連続DPについて紹介し，さらに本論文で新たに提案する認識手法を概説する．

1.3 連続DP

動的計画法（DP）は，1957年にR.Bellmanが定式化したもので，順序関係のない対象物についての「DP」であるため，順序の逆転も考えられていた．しかし，順序関係が その物理的な性質からの本質的な拘束となっている音声やジェスチャの時系列データを扱うに至って，順序の逆転の無い単調増加なパスのみが DP の適用において取られるようになった（これをDTW(Dynamic Time Warping)と呼ぶこともある）．この拘束は,「傾斜制限」の名で呼ばれ，DPが音声認識において多用されるきっかけとなった．

しかし，このDPでは入力パターンの事前の切り出しが必要であるために，連続音声の認識が困難であった．このような状況で，岡は「連続DP」[15]という入力パターンの切り出しとマッチングをフレームワイズに同時的に実現する（スポッティング認識可能な）手法を提案し，音声認識やジェスチャ認識にて活用されるようになった．

さらに，伊藤ら[60]は，参照パターンの切り出しをも同時に行なう「Reference Interval-Free 連続DP(RIFCDP)」を提案した．この手法によって時系列パターン同士の任意の区間長の類似区間を検出できるようになった．

しかし，RIFCDPの計算量とメモリ量は膨大，戸惑い動作の認識は困難という課題が残っていた．そこで，本論文ではこの課題の解決を目指し，局所距離にかかる重みを，標準パターンにおいて過去に溯るに従って指数関数的に減少させて累積距離を計算する「重み減衰型RIFCDP」と呼ぶ手法を提案する．この手法は，RIFCDPと比較して計算量とメモリ量を約1桁削減できるという特長がある．

また，局所距離にかかる重みを，入力軸方向において過去に溯るに従って指数関数的に減少させて累積距離を計算する「非単調連続DP」と呼ぶ手法も提案する．これは，連続DPで必要となっていた単調増加(monotonic)な傾斜制限を無くし，右肩下がりのパスをも許すものもである．従って，静止したり逆に戻ったりする戸惑い動作も認識できる．また，この手法によって，ロボットが経路中を行きつ戻りつ移動しても位置推定が可能となる．

「重み減衰型RIFCDP」と「非単調連続DP」とでは，連続DPにおいて累積距離を求める際に，いずれも局所距離にかかる重みを指数関数的に減少させている．この減少させる方向が，標準パターンの軸方向か入力軸方向かが異なる．

次節以降にて，時系列画像認識において時系列パターンを分類することが必要となる以下の3分野への適用を図る．

[1] 動画像を用いた人物のジェスチャ認識

[2] 周辺環境の時系列画像を用いた移動ロボットの自己位置推定

[3] 時系列データからの類似時系列の検出
これらは，いすれも時系列画像を利用しており，各フレームから得られる特徴ベクトルは，時系列データである．

1.4 動画像を用いたジェスチャ認識

人間の身振り手振りを認識する技術は，柔軟な Man-Machine Interface System を構築する上で重要である [16][18]．特に，動作者にデータグローブ等の接触型センサやマーカーを装着させることなく，人間の動作を捉えた動画像からジェスチャ認識が可能となれば，より自然なインターフェースを実現できる [62]．さらに，これによって，キーボードやマウスの操作に煩わされることなく自由に動き回りながら自然な動作で指示を出すことも可能となる．また，動画像よる手話認識が可能となれば，データグローブなどを必要とするシステムと比べて衛生的で気軽に使えるものなるだろう．

画像からの特徴抽出法としては，対象を 3 次元形状モデルで表現し，入力画像に当てはめて各種パラメータを抽出する手法 (3 次元モデルに基づく手法)[19]～[24] と，見かけの変化から特徴を抽出する手法 (見え方に基づく手法) [25]～[29] とに大別できる．後者の手法の方が人物や背景に対する制限が厳しいものの，計算量が少なく実時間認識システムを実現しやすい．すでに，高橋ら [29] は，見え方に基づく特徴抽出法を用いたジェスチャのスポッティング認識手法を提案している．しかし，この提案には衣服と背景の明るさの変化や動作軌跡の変動に弱いなどの問題があった．そこで，本論文では，この特徴抽出における問題を解決するために新たな動きと形状特徴を提案し実時間認識システムを実現して本手法の有効性を実証する．さらに，高橋らの認識システムでは，新たなモデルをオンラインで教示することは困難という問題があった．そこで，本論文では，オンライン教示を実現する．

また，人間のジェスチャは，同一動作であっても途中で戸惑ったり考えて止まったりすることがある．例えば，教室で生徒が手を挙げるとき，自信たっぷりに素早く挙げるときもあれば若干躊躇ながら挙げるときもあろう．このように，同じ動作でも人物の気持ちによって変化することが想像される．従って，このようなジェスチャの認識が可能となれば，その戸惑い具合から動作者の本音やその動作の自信のほどを読み取ることができ，コンピュータエージェントがより繊細な応対を行なうことができるだろう．

しかし，この戸惑っている動作パターンは，時と場合によって無数に変化すると考えられる．そのため，標準パターン全体との距離を求める単調増加な傾斜制限をもつ従来の連続 DP では，戸惑いパターンに対応した多くの標準パターンを用意する必要が生じ非効率的となっていた．そこで，本論文では，この認識手法における限界を克服すべく，非単調連続 DP を提案する．

1.5 時系列画像を用いた移動ロボットの自己位置推定

オフィス，家庭のような実世界における移動ロボットの位置推定，誘導法では，実時間処理が必要であるとともに，スリップや揺れなどの走行外乱に対するロバスト性が要求される．このような場合，内界センサのみによる位置推定法では誤差が蓄積されやすい．そこで，外界センサから環境情報を随時取得し，あらかじめ獲得した環境モデルと照合して蓄積された誤差を低減する方法が用いられている．環境に人工的な目印を設置しない位置修正方法としては，Kalman フィルタを用いて超音波センサの補正を行い現在位置を推定する手法[64]やセンサの信頼性を考慮してランドマークを決定する手法[65]などが提案されている．また，Kalmanフィルタで複数のセンサ情報を統合する手法[66]も提案されている．

一方，視覚センサを用いれば，人が周辺の風景から自己の位置が分かるのと同様に，有用な位置固有の情報が得られると考えられる．視覚センサによる位置推定法には，センサ情報から獲得した環境の幾何モデルを環境モデルとする方法(モデルに基づく手法)とセンサ情報を単純に加工した情報を環境モデルとする方法(センサーに基づく手法)とがある[68]．前者の方法では，センサ情報から幾何モデルを獲得する過程で失敗すると位置推定が困難となる．また，画像情報から幾何モデルを獲得するときに，多くの計算量を必要とする．これに対し，後者のセンサ情報を用いる方法では，センサ情報を直接もしくは単純に融合・圧縮して環境モデルを作成する．このため，単調な景色が連続している環境や類似した景色が複数箇所にある環境では，位置推定が困難となるものの，位置固有の有用な情報が得られれば前者の方法よりロバストな方法となり得る．

そこで，環境モデルとして図1.1に示すような環境を撮影した画像系列からなるトポロジー地図を用い，移動ロボットの位置推定を行う[75]．ここで目標とした機能は以下のようである．まず，地図作成時に「ここは，Aさんの席」「ここは，本棚の前」などの情報を画像のフレーム番号に付随して記憶しておく．自律走行時は，得られる画像と地図内の画像系列とを絶えずマッチングして自己位置を推定している．例えば「本棚の前まで案内して下さい」という指示が与えられると，現在地点から目標地点までの画像系列（経路）を決定し，この画像を順次たどるように移動する．

このとき，超音波センサと赤外線センサにより障害物回避を行うものとし[74]，環境照明の変化が小さいものとしている．さらに，位置推定により得られるのは，地図中の画像系列の中で最もマッチングした画像の番号だけである．つまり，地図中の大体の位置（大局的位置）を推定するのであり，地図画像取得位置に対する正確な自己位置や方向（局所位置）は得られない．本論文では，大局的位置の推定により次のような誘導を行うことを目的としている．つまり，「通常は障害物回避を行いながら進み，誘導経路である画像系列とマッチングできなくなったら誘導経路から外れたものとして反転して戻り，別の道を選択する」ということを繰り返して，最終的には目的地に到達するという誘導である．

地図画像列を用いた位置推定では，はじめに画像からの特徴抽出を行ない，次に入力特徴

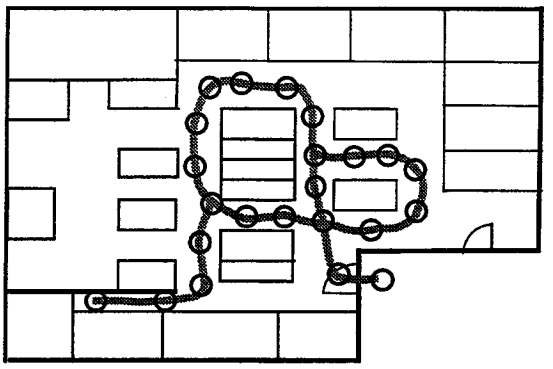


図 1.1 トポロジー地図

とマッチングして地図中の位置を推定する．我々は，既にロボット走行時のカメラの揺れにロバストになるよう，複数連続した低解像度な画像特徴を用いてマッチングする手法を提案している [75]．この手法では，単調な景色が連続した環境には適用が困難であるが，Reference Interval-Free 連続DP(Dynamic Programming) [60] を用いてスポッティング的な位置推定を可能とした．しかし，移動ロボットの移動方向が地図作成時と同じ方向に制限されていた．そこで，本論文では，ロボットが地図作成時と逆方向に走行したときや停止時でも同じ枠組で位置推定を可能とするために，すでにジェスチャ認識のために提案した非単調連続DPを用いることを提案する．また，ロボットの向きに関わらず大局的位置の推定を可能とするために有用である回転不変な特徴量として，全方位視野を水平に分割し，各領域内の局所的な特徴の平均値を用いる．

1.6 時系列データからの類似時系列の検出

近年，膨大なマルチメディアデータの検索や要約，モーダル変換，構造解析を実現できる時系列パターンの検索技術が，重要な研究課題となっている．特に計算機の計算能力の向上や記憶容量の増大とともに，ディジタル化された映像メディアのハンドリング技術に関心が集まっている．実際，数十年分のテレビ映像を個々人で持ち必要な部分を自由に検索したり編集することも夢ではなくなってきた．また，人間が目で捉えた映像を対象としてこのようなシステムを実現すれば，記憶の補助に役立つだろう．また，部屋や町の複数箇所に設置したカメラからの映像を活用して交通状況の把握，犯罪の防止や操作の支援などを行うことも考えられる．

現状の映像ハンドリング技術 [58] では，効率良く必要な情報を検索する方法と多量の映像データを分かりやすく表現する方法が研究されている．検索のためには，画像からの特徴抽出，移動体などの追跡，映像に付随した文字情報の抽出などを行い，映像情報の構造化およびインデキシングを行うことことが必要である．しかし，時系列画像を扱っているにもかかわらず，従来では時系列データの変化パターンを利用した検索は行われていなかった．そこで，本論文で

は,有効な特徴パラメータが抽出されたことを前提とし，この時系列データの中から変化パターンの特徴を抽出することを目指す．つまり，２つの時系列データ間について，任意の長さを持ち，かつ互いに類似した区間（類似区間）対を検出することを目標とする．

このような類似区間の検出手法は，音声認識の分野で提案されている．兵後らは，始終端固定のDPによる発話単位の類似区間を検出する方法を提案している[59]．しかし，この手法では，複数の類似区間がある場合，その出現順序が２つの時系列パターン間で同じでなければ検出できなかった．そこで，伊藤らは，Reference Interval-free 連続DP(RIFCDP) [60] を提案して，切り出しや認識を行なっていない音声などの時系列の事例データベースから直接検索を行えることを示した．この，RIFCDPでは，類似区間の順序，個数の制限が無い．

しかし，従来のRIFCDPでは標準パターン中の各フレームにおいて，検出したい類似区間長分の累積距離および入力時刻を保持・算出しているため，メモリ量と計算量が非常に大きくなる．

そこで，本論文では，1.3節にて述べたように局所距離にかかる重みを過去に溯るに従って指数関数的に減少させて累積距離を計算する考えを導入することにより，計算量とメモリ量の軽減を可能とし，かつほぼ類似の機能をもつ 重み減衰型 RIFCDP を提案する．本手法を用いると，従来のRIFCDPの約50分の１のメモリ量でシステムが実現できる．また，計算量についても従来の約1/16となる．また，これまでRIFCDPが適用されていなかったジェスチャ動画像[29] を用いて本手法の有用性を示す．これにより，今まで音声の分野で主に用いられていた検索や要約，話題境界検出技術を，人物の行動理解，手話認識，さらにはテレビ映像などのマルチメディアを扱う分野へ適用できることを示す．

1.7 本論文の構成

本論文では，第２章 にてジェスチャ認識における新たな特徴抽出法として，動きと形状特徴を提案し，さらに第３章ではオンライン教示システムを実現する．第４章では,非単調連続DPを提案し，戸惑い動作の認識や移動ロボットの自己位置推定へ適用する．また，第５章では,重み減衰型RIFCDPを提案し，ジェスチャ動画像の検索を行ってその有効性を示す．

第 2 章

動きと形状特徴を用いた連続DPによるジェスチャ認識

2.1 はじめに

スポッティング認識可能な連続DPを用いることで動作者にジェスチャの開始と終了を意識させないジェスチャ認識システムが，高橋らによってすでに提案されている[29]．この研究では，以下の条件を設定して白黒動画像から見かけベースな動き特徴を抽出している．

[条件1] 画像中には一人の人物が一定位置，一定の向きで存在し，背景の時間的な変化は少ない．

[条件2] 認識対象は，動作の大きなジェスチャとする．(手の形などの細かい動きは認識対象としない.)

具体的には，図2.1に示すように，はじめにサイズが64 × 64の入力画像の時間差分画像を求める．次に，平均化処理により画像サイズを16 × 16に縮小し，時間方向に3フレーム分平均化する．最後に，すべてのピクセル値の対数を求めて，これを16 × 16次元の特徴ベクトルとして連続DPへの入力とする．つまり，カメラの視野を16 × 16に分割し，各小領域中の画素値の時間変化の平均を特徴ベクトルとしている．

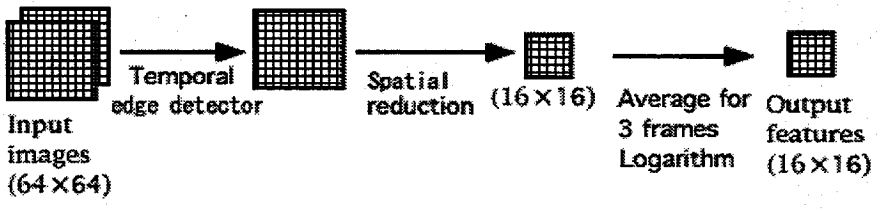


図 2.1 従来の特徴抽出法

しかし，この特徴抽出法において以下の問題があった．

[問題1] 時間差分画像の濃淡値を用いるため，衣服と背景の明るさの変化に弱い．

[問題2] カメラの視野を16 × 16と細かく分割し過ぎているため，動作軌跡の変動に弱い．

[問題３] 手の形状や方向（手形状）が異なっていても，類似した動きであれば判別が困難．時間差分画像の濃淡値は，人物と背景の明るさの違いである．従って，この濃淡値をそのまま用いる従来法では，原理的に人物と背景の明るさの変化によって大きな影響を受ける．高橋らの報告[29]では衣服と背景を変化させて評価実験を行っているが，明るさの変化が小さかったために［問題１］は生じていなかったと思われる．

また，日常用いられる人物動作では，多少の軌跡変動が避けられない．従って，従来法のようにカメラの視野の分割数が大きく，１つの小領域あたりの視野が狭いと，同一のジェスチャでも視野全体に手の一部が通過する場合と全く通過しない場合とが生じる．従来の特徴抽出法では特徴ベクトル間の距離演算において空間方向の変化を吸収していない．従って，従来法では，若干の動作軌跡の変動が特徴値を大きく変化させ［問題２］を生じる．

さらに，通常のジェスチャでは，手の形状が重要な役割を果たす．しかし，従来の特徴抽出法では，画像中に含まれる腕，体，頭などと手の部分との区別を行っていない．このため，手形状に関する情報のみを得ることができずに［問題３］が生じる．

そこで，これらの問題を解決するため，新たに動きの特徴と手の形状特徴の抽出法を提案する．この手法では，２値化した時間差分画像を用いて画像中の変化領域を求めることで［問題1］で述べた衣服・背景の変化に対処する．また，動き特徴においては，低解像度の画像特徴を用いることで動作変動を吸収し，[問題2] を解決する．さらに，形状特徴においては，時間差分画像の２値化により抽出した領域がおよその手形状を保存するため，この抽出領域の輪郭の方向の分布を特徴とする．ただし，このとき，手の大きさや腕，体，頭との位置関係から推定した重みをかけて方向分布求める．これによって，主に手部分に関する輪郭方向の分布を求め，[問題3] で述べた手形状を判別できないという問題を低減する．

本章の構成は，2.2節にて動きと形状特徴を提案し，2.3節にて従来から用いられている連続DPによるスポッティング認識について説明する．また，2.4節にて動き特徴を評価し，2.5節にて動きと形状特徴を評価し，2.6節でまとめる．

2.2 動きと形状特徴

2.2.1 動きと形状特徴の概要

本節では，従来と同様に白黒動画像から得られる見かけベースな特徴を用いるという前提で，手の動きと形状特徴を抽出する，図2.2のような特徴抽出方法を提案する．ここでは，時間差分２値画像から動き特徴値（$f_m(k,t)$と記す，図2.2左方）とともに形状特徴値（$f_s(k,t)$と記す，図2.2右方）をも抽出する．この時間差分２値画像がおよその手形状を保存することは2.2.4節にて示す．

動き特徴は，画像をいくつか（図では3×3）に分割し，各領域中で時間的に変化した割合を求めたものである．また，形状特徴は，輪郭を検出しそのチェインコード（輪郭の移動方向

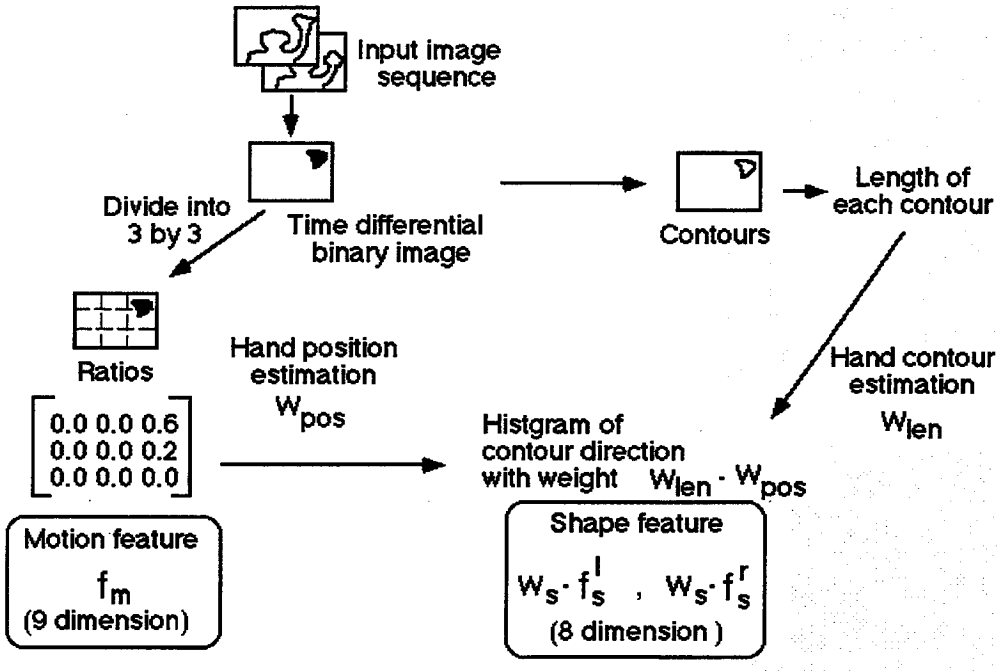


図 2.2 動き特徴と形状特徴の抽出

列）から4方向の出現頻度を求めたものである．このとき，左右の手に関する情報を得るために画像を左右に分割している．さらに，この2つの特徴値の絶対値が異なるため，重みw_sを形状特徴値にかけて作成した特徴ベクトル（図では17次元となる）が，ここで提案する動きと形状特徴となる．

2.2.2 動き特徴

本節では，動き特徴（図2.2左方）の定式化を行う．初めに，サイズ$N_1 \times N_1$の入力画像$I(i,j,t)(0 \leq i,j < N_1, 0 \leq t)$と1フレーム前の入力画像$I(i,j,t-1)$から，次式により2値画像$I_b(i,j,t)$を求める．

$$I_b(i,j,t) = \begin{cases} 1 & \text{if } |I(i,j,t) - I(i,j,t-1)| \geq h_c \\ 0 & \text{otherwise.} \end{cases} \tag{2.1}$$

ただし，h_cは画素値が変化したか決定するしきい値である．さらに，次式のように2値画像$I_b(i,j,t)$をサイズ$N_2 \times N_2$に縮小し，特徴ベクトル$f(k,l,t)(0 \leq k,l < N_2)$を求める．

$$f(k,l,t) = \frac{1}{h^2} \sum_{0 \leq p,q < h} I_b(k \cdot h + p, l \cdot h + q, t). \tag{2.2}$$

ここで，pとqはともに整数，$h = N_1/N_2$である．この特徴ベクトル$f(k,l,t)$は，入力画像を$N_2 \times N_2$に分割した各領域内において画素値が変化した割合，つまり，対象が移動したと推定される領域の割合である．また，特徴ベクトル$f(k,l,t)$の各要素はh^2の分解能を持つため，N_1/N_2

が小さくなると量子化誤差が大きくなる．さらに，従来法の対数を取る代わりにあるしきい値以上を飽和させる．

$$f'(k,l,t) = \min\{f(k,l,t), h_m\}. \tag{2.3}$$

この処理は対数計算と比較して計算量が小さいだけでなく，値が小さいときの雑音を強調しないという利点があると考えられる．

まとめると2.1節で述べた［問題１］は，式(2.1)の２値化処理によって変化領域を求めることで解決する．従来法で用いていた背景と衣服との濃淡値の差に比べ，pixelごとに変化したか否かを決定し，3×3の各領域内での変化したpixel数の割合を求めるため衣服と背景の変化の影響を受けにくい．また，［問題２］は特徴ベクトルの次元$N_2 \times N_2$を3×3程度に小さくすることで解決する．これによって，手の移動軌跡が若干変化しても特徴ベクトルの変化が小さく，ある程度以上変化してはじめて特徴値が大きく変化する．このことを図2.2に図示した．図2.2(a)は，理想的な特徴値の変化を示し，図2.2(b)は，従来法における変化を示した．従来法では，軌跡の変化と特徴値の変化が比例している．これに対して，図2.2(c)に示した提案手法では，理想的な変化に近いことが分かる．

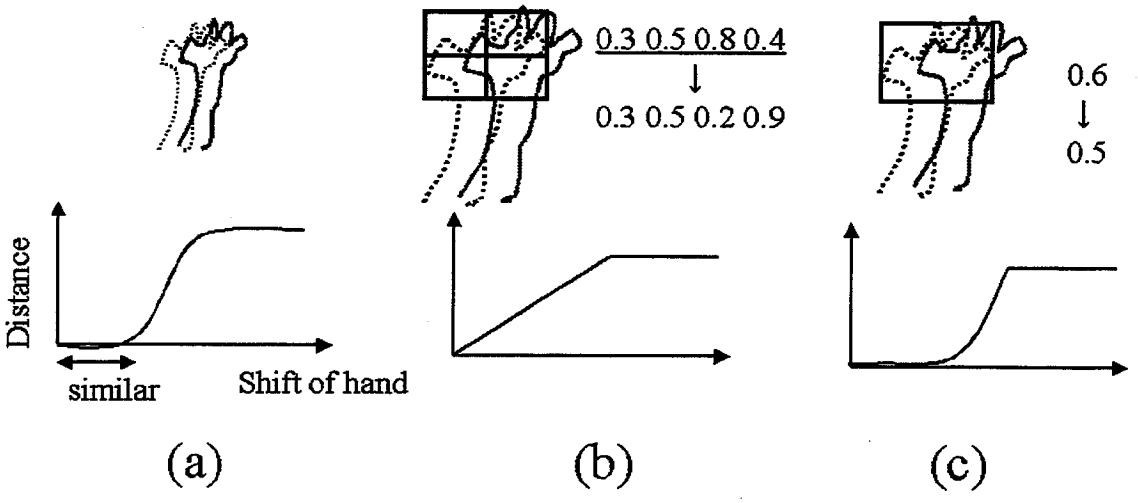


図 2.3 軌跡変化時における特徴ベクトルの変化
(a)理想的な特徴変化:微小な軌跡変化時は特徴変化小，大きな軌跡変化で特徴変化大，(b)従来法，(c)提案手法

このとき，入力画像のサイズ$N_1 \times N_1$が12×12程度でも高い認識率を実現できることを2.4節の評価実験で示す．この入力画像の縮小化により，低解像度の人物のジェスチャ認識が可能となる．また，得られる特徴ベクトルの次元数の低下により連続DPの計算量を一桁程度低減できる．

2.2.3 形状特徴

本節では，形状特徴（図2.2右方）について述べる．はじめに，時間差分2値画像から閉領域の輪郭をチェインコードとして抽出する．このチェインコードの値つまり，輪郭方向値$cc(p,q,t)$ $(p=1,\cdots,P(t), q=1,\cdots,Q(p,t))$は，図2.4に示すように8方向を表す0〜7の整数で記述する．ただし，$P(t)$は時刻tにおける画像中の検出閉領域の総数，$Q(p,t)$は各閉領域の輪郭の長さとする．次に，図2.4のように8方向を4方向に統合させ，この各方向が画像中で出現する頻度$\mathrm{hist}(k,t)(k=0,\cdots,3)$を求める．

$$\mathrm{hist}(k,t)=\sum_{p=1}^{P(t)}\sum_{q=1}^{Q(p,t)}\{\delta(cc(p,q,t),k)+\delta(cc(p,q,t),k+4)\}\quad(k=0,\cdots,3) \tag{2.4}$$

ただし，関数$\delta(a,b)$は，$a=b$のときのみ1，他の場合は0とする．

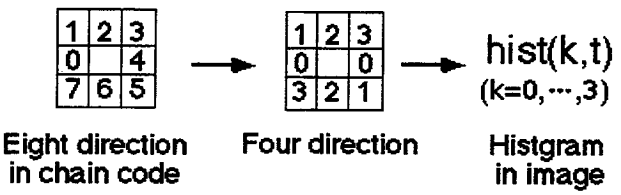


図 2.4 チェインコードからの方向分布抽出

しかし，時間差分2値画像には腕，体，頭などの手以外の動きによる領域が雑音となる．そこで，局所的な輪郭方向に対して，手の大きさや位置関係などの大局的な情報から推定した重みをかけて方向分布を求めることで手以外の影響を減少させる．本章で提案する形状特徴の値$f_s(k,t)$は次式のように記述できる．

$$f_s(k,t)=\sum_{p=1}^{P(t)}w_{\mathrm{len}}(l(p,t))\cdot\sum_{q=1}^{Q(p,t)}w_{\mathrm{pos}}(i(p,q),j(p,q),t)\cdot \{\delta(cc(p,q,t),k)+\delta(cc(p,q,t),k+4)\}\quad(k=0,\cdots,3) \tag{2.5}$$

但し，入力の時刻tにおける時間差分2値画像中のp番目の領域の輪郭の長さを$l(p,t)$とし，重み$w_{\mathrm{len}}(l)$ $(0\le w_{\mathrm{len}}(l)\le 1)$は，抽出領域の輪郭長が$l$のときに，手指であると判断できるほど大きい値となるように設定する．また，座標$(i(p,q),j(p,q))$はチェインコード$cc(p,q,t)$が得られた画像上の座標であり，重み$w_{\mathrm{pos}}(i,j,t)$ $(0\le w_{\mathrm{pos}}(i,j,t)\le 1)$は，動き特徴値$f_m(k,l,t)$から画像中の手の位置を大局的に推定して，座標$(i,j)$が手の部分であると判断できるほど大きい値とする．この二つの重みに関しては，2.2.5節,2.2.6節にて説明する．

さらに，両手が画像中の左右に存在することが多いことを考慮して画像を左右に二分し，各領域中の輪郭の方向分布を各領域面積で割ったものを$f_s^l(k,t), f_s^r(k,t)$とする．従って，本章で

提案する特徴値は，以下のように動き特徴と形状特徴とを合わせた17次元のベクトルとなる．

$$\{f_m(0,0,t),\cdots,f_m(2,2,t),w_s\cdot f_s^l(0,t),\cdots,w_s\cdot f_s^l(3,t),\\ w_s\cdot f_s^r(0,t),\cdots,w_s\cdot f_s^r(3,t)\} \tag{2.6}$$

ただし，重みw_sは，形状特徴に対する重みであり，$w_s=0$のときは動き特徴のみとなり，w_sが大きいほど形状特徴の寄与率が上昇する．この重みw_sについては，2.5節にて最適化を行なう．連続DPによる認識では，ここで得られた17次元$(N=17)$の特徴ベクトルを用いて式(2.15)を算出し，局所距離$d(t,\tau)$を求める．

2.2.4 手領域の抽出

2.2.2節で述べた低解像度特徴は，動き情報の抽出に主眼が置かれており，式(2.1)による時間差分画像では手形状の抽出が困難である．そこで，時間差分2値画像$I_2(i,j,t)$を以下の式で求める．

$$I_2(i,j,t)=\begin{cases}1 & \text{if } \min_{|k|\le 1,|l|\le 1}\{I_d(i,j,t,k,l)\}\ge h_c\\ 0 & \text{otherwise.}\end{cases} \tag{2.7}$$

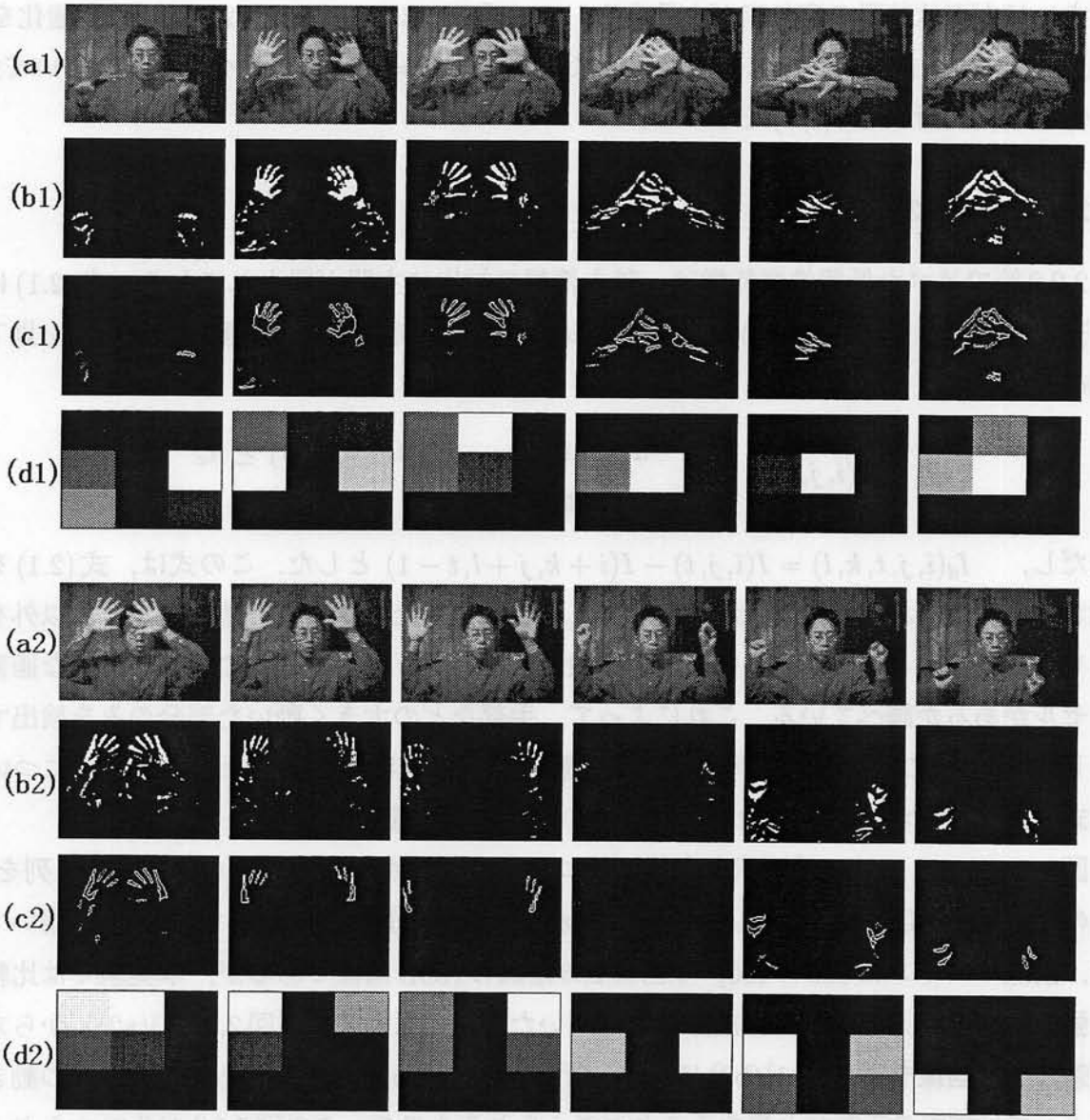
ただし，　$I_d(i,j,t,k,l)=I(i,j,t)-I(i+k,j+l,t-1)$ とした．この式は，式(2.1)を若干変化させたものであり，画素値がしきい値h_c以上増加したピクセルの値を1，それ以外を0とするものである．このとき，小さな動きを検出しないように，対応点の周囲に同様な画素値のピクセルがあるか調べている．これによって，手部などの大きく動いた部分のみを検出するようにした．2.5節の動きと形状特徴の評価実験では，形状特徴だけでなく動き特徴についてもこの式(2.7)で求めた時間差分2値画像を用いて特徴抽出を行う．

図2.5(a1)(a2)には，手形状を変化させつつ両手を動かす動作を行なった画像系列を示した．この画像は，SGI社のIndy(R4400 200MHz)と，付属のIndyComというカメラを用いて作成した．画像のサイズは160×120，1画素256階調のRGB画像であるが，本実験では比較的輝度に強い影響を与えるGREEN成分のみを用いた．この原画像列（図2.5(a1)(a2)）から求めた時間差分2値画像を図2.5(b1)(b2)に示す．背景が手の明るさと同等である場合や手の動きの方向によっては実際の手形状と異なることが考えられるものの，この図2.5(b1)(b2)からおよその手の形状や方向を見てとれることが分かる．

しかし，手を画像中の上の方で動かすジェスチャの場合や腕や体も動かすジェスチャの場合には，手以外の部分が検出され，これが手の形状を求める上で雑音となる．そこで，2.2.5節,2.2.6節において手の特性を用いた重み推定を行ない，手以外の雑音を低減する．

2.2.5 輪郭長による重み

本節では，式(2.5)で導入した輪郭長による重み$w_{\mathrm{len}}(l)$について述べる．はじめに，対象とするジェスチャ動画像の時間差分2値画像について調べ，輪郭の長さが$l(0\le l\le L)$である領



(a1),(a2):原画像（8フレーム毎），(b1),(b2):時間差分2値画像, (c1),(c2):輪郭長から推定された重みw_{len}, (d1),(d2):動き特徴から推定された画像中の位置による重みw_{pos}

図 2.5 輪郭方向に対する2種類の重み（白：1，黒：0）

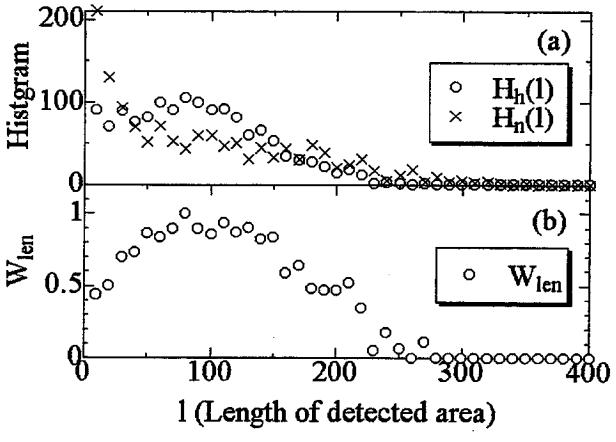
域が手であった場合は$H_h(l)$に，そうでなかった場合は$H_n(l)$に投票する．次に，以下の式で$w_{\mathrm{len}}(l)$を求める．

$$w_{\mathrm{len}}(l) = \frac{w'_{\mathrm{len}}(l)}{\max_{0\le l\le L} w'_{\mathrm{len}}(l)}. \tag{2.8}$$

$$w'_{\mathrm{len}}(l) = \frac{H_h(l)}{H_h(l) + H_n(l) + \sigma_l}. \tag{2.9}$$

ただし，σ_lは，該当する輪郭長に対する投票数が小さく信頼性が低い場合に$w_{\mathrm{len}}(l)$を減少させるための定数である．

今回の実験では，図2.18に示すような評価実験で用いたジェスチャ動画像を3フレーム毎に調べ，$w_{\mathrm{len}}(l)$を求めた．ただし，輪郭の長さlは$0 \le l \le 400$の範囲について5ステップで分割し投票した．ここで用いた画像数は221枚であり，また投票数を調べて$\sigma_l = 10$とした．図2.6(a)には，$H_h(l)$と$H_n(l)$を，図2.6(b)には，$w_{\mathrm{len}}(l)$を示した．



(a) 投票数$H_h(l)$と$H_n(l)$，(b) 重み$w_{\mathrm{len}}(l)$.

図 2.6 重み$w_{\mathrm{len}}(l)$の算出結果

図2.5(c1),(c2)には各輪郭を重み$w_{\mathrm{len}}(l)$が0のときに黒，1のときに白に段階的に対応させて表示した．この図から，手部分と推測し難い極端に小さな領域の重みが小さくなっていることが分かる．

2.2.6 手の位置による重み

本節では，式(2.5)で導入した重み$w_{\mathrm{pos}}(i,j,t)$について述べる．重み$w_{\mathrm{pos}}(i,j,t)$は，画像中の座標(i,j)に手が存在する確率を示し，動き特徴$f_m(k,l,t)$から大局的に推定する．2.2.2節と同様に，入力画像サイズを$N_1 \times N_1$，特徴の次元を$N_2 \times N_2 (N_2 = 3)$として，

$$w_{\mathrm{pos}}(i,j,t) = \frac{w_0(i_h,j_h)\cdot f_m(i_h,j_h,t)}{\max_{0\le k,l<N_2} w_0(k,l)\cdot f_m(k,l,t)} \quad (0 \le i,j < N_1) \tag{2.10}$$

とする．ただし，i_h, j_hは，それぞれi, jを$h(h = N_1/N_2)$で割った商の整数部分，$w_0(k,l)(0 \leq k,l < N_2, 0 \leq w_0(k,l) \leq 1)$は，補正重みである．この補正重み$w_0(k,l)$は，以下のように決定する．はじめに，対象とするジェスチャ動画像の時間差分2値画像について，画像を$N_2 \times N_2$分割し，$(k,l)(0 \leq k,l < N_2)$番目の分割領域に手が存在した場合に$H_h(k,l)$に$f_m(k,l)$を加え，そうでなかった場合には$H_n(k,l)$に$f_m(k,l)$を加える．これをすべての画像について各分割領域毎に行う．つぎに，以下の式で$w_0(k,l)$を求める．

$$w_0(k,l) = \frac{w_0'(k,l)}{\max_{0 \leq k,l < N_2} w_0'(k,l)}. \tag{2.11}$$

$$w_0'(k,l) = \frac{H_h(k,l)}{H_h(k,l) + H_n(k,l) + \sigma_p}. \tag{2.12}$$

ただし，σ_pは，$H_h(k,l)+H_n(k,l)$が小さく信頼性が低い場合に$w_0(k,l)$を減少させるための定数である．2.2.5節と同様に図2.18で示したジェスチャ動画像について調べた．$H_h(k,l)+H_n(k,l)$の値から$\sigma_p = 5$とし，表2.1に示すように$w_0(k,l)$を求めた．kが大きいほど小さな値になっている．これによって，例えば手部と腕部が同時に存在した場合は通常上の方に存在する手部の方が強調される．図2.5(d1),(d2)には3×3の動き特徴から得られた重み$w_{\mathrm{pos}}(i,j,t)$を0が黒，1が白に対応させて段階的に表示した．この図から，およその手の位置が推定できていることが分かる．

表 2.1 手の位置による重みを求めるための補正重み$w_0(k,l)$
カッコ内は左から，$H_h(k,l), H_n(k,l)$の整数部．

	$l=0$	$l=1$	$l=2$
$k=0$	1.00(24,5)	0.85(15,5)	0.93(21,6)
$k=1$	0.69(45,37)	0.72(52,45)	0.66(41,42)
$k=2$	0.32(32,98)	0.32(21,67)	0.37(35,93)

2.3 連続DPによるスポッティング認識

一つのモデルZを，標準動作を捉えたTフレームの動画像から得られる特徴ベクトルz_τの系列（これを標準パターンと呼ぶ）

$$Z = \{z_\tau | 1 \leq \tau \leq T\} \tag{2.13}$$

で表す．ここで，特徴ベクトルz_τはその次元数をNとすると

$$z_\tau = (z_\tau(1), z_\tau(2), \cdots, z_\tau(N)) \tag{2.14}$$

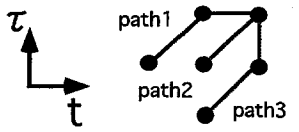


図 2.7 連続DPの局所パス

である．入力画像からも同様な特徴ベクトル系列$u_t(0 \leq t < \infty)$が連続的に得られる．このとき，u_tとz_τとの局所距離を$d(t,\tau)$と表記する．本章で用いた$d(t,\tau)$の定義を以下に示す．

$$d(t,\tau) = \frac{1}{N}\sum_{k=1}^{N}(u_t(k) - z_\tau(k))^2. \tag{2.15}$$

ここで，入力，モデルの時間軸をそれぞれt,τと区別する．

さらに，点(t,τ)を終点としたモデルと入力系列との累積距離を$S(t,\tau)$で表す．連続DPでは$S(t,\tau)$を以下のような漸化式で更新する．

初期条件$(t=0)$：

$$S(-1,\tau) = S(0,\tau) = \infty. \ (1 \leq \tau \leq T) \tag{2.16}$$

漸化式$(1 \leq t)$:

$$S(t,1) \ = \ 3 \cdot d(t,1). \tag{2.17}$$

$$S(t,2) = \min \begin{cases} S(t-2,1) + 2 \cdot d(t-1,2) + d(t,2) \\ S(t-1,1) + 3 \cdot d(t,2) \\ S(t,1) + 3 \cdot d(t,2). \end{cases} \tag{2.18}$$

$$S(t,\tau) = \min \begin{cases} S(t-2,\tau-1) + 2 \cdot d(t-1,\tau) + d(t,\tau) \\ S(t-1,\tau-1) + 3 \cdot d(t,\tau) \\ S(t-1,\tau-2) + 3 \cdot d(t,\tau-1) + 3 \cdot d(t,\tau). \end{cases} \tag{2.19}$$
$$(3 \leq \tau \leq T)$$

ここで，tは入力の離散時刻を表し，τは標準パターンの長さに対応するパラメータで，$1 \leq \tau \leq T$ (Tは標準パターン長) である．この漸化式では，図5.4に示す3個の局所パスのうちで累積距離が最小となる値が選択されている．このため，標準パターン全体との累積距離$S(t,T)$は，入力の時間方向の伸縮が$\frac{1}{2}$〜2倍であるとしたときの最小の累積距離となっている．連続DPの出力$A(t)$は，重みの和$3 \cdot T$で正規化して$A(t) = \frac{1}{3 \cdot T}\ S(t,T)$と定める．

ここで，モデルがL個存在するとし，各パターンの累積距離を$A_\ell(t)(1 \leq \ell \leq L)$，しきい値を$h_\ell$とする．認識結果は，マッチングしたモデルのカテゴリー番号$\ell^*(t)$であり，以下の式で

判定する．

$$\ell^*(t) = \begin{cases} \mathrm{Arg}[\min_\ell\{A_\ell(t) - h_\ell\}] & \text{if } \exists\ell \quad \text{so that} \quad A_\ell(t) \leq h_\ell \\ \mathrm{null} & \text{otherwise} \end{cases} \tag{2.20}$$

ここで，Argは引数ℓを返す関数，nullは空のカテゴリーを表す．

2.4 動き特徴の評価

2.4.1 実験方法

実験装置として，SGI社のIndy(R4400 200MHz)と，付属のIndyComというカメラを用いた．実験は，オフィス内で椅子に座った1人の被験者に対して行った．カメラの視野は被験者のジェスチャが適切に入るように設定した．また，照明は建物の天井に設置されている蛍光灯のみを用いた．

CCDカメラの出力映像をAD変換して得られる画像は，サイズ160×120，1画素256階調のRGB画像であるが，認識には比較的輝度に強い影響を与えるグリーン成分のみを用いた．実験では，この入力画像のサイズを$N_1 \times N_1$に縮小し，特徴抽出部への入力とした．また，式(2.3)の特徴ベクトルの値を飽和させるしきい値h_mは0.3とした．

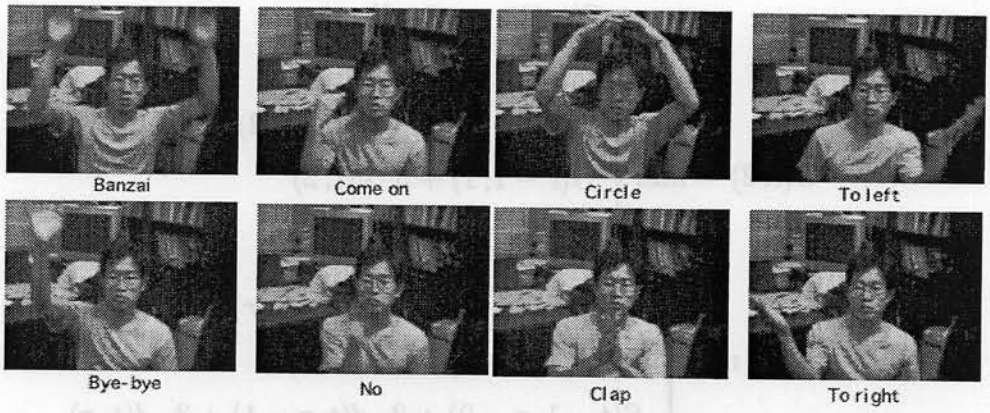


図 2.8 8種類のジェスチャのスナップショット

実験に用いたジェスチャは，(1)ばんざい（両手），(2)バイバイ（右手），(3)まる（両手），(4)手をたたく（両手），(5)こちらへ（右手），(6)左へ（左手），(7)右へ（右手），(8)いいえ（右手）の8種類である．これを，ジェスチャ$\ell(\ell = 1, 2, \cdots, 8)$と表記する．図2.8に各ジェスチャのスナップショット，図2.9にジェスチャ“バンザイ”の画像系列を示す．被験者は各動作を通常のスピードで行い，画像は15Hzでサンプリングした．また，式(2.1)の閾値h_cはカメラの熱雑音を考慮し10とした．

標準パターン$\ell(\ell = 1, 2, \cdots, 8)$は，それぞれのジェスチャを捉えた画像系列から人手でジェスチャ部分のみを切り出し作成した．この実験で用いた標準パターンのフレーム長Tは11から

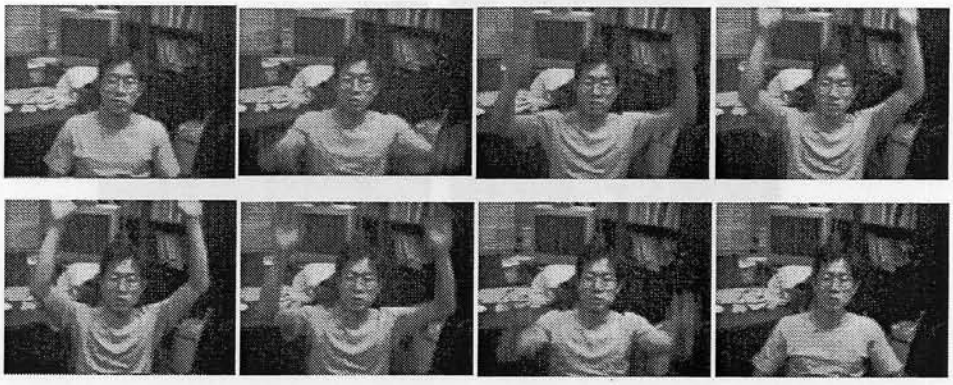


図 2.9 ジェスチャ“ばんざい”(2フレーム毎)

15であった．また，同じジェスチャを20回繰り返した入力画像列ℓを作成した．次に，入力画像列ℓを認識システムに入力し，1位認識率と正解候補率を求めた．

$$AFC(\text{1位認識率}) = \frac{\sum_{\ell=1}^{8} \text{正答ジェスチャ数}\ell}{\text{全ジェスチャ数}(20 \times 8)} \tag{2.21}$$

$$RCC(\text{正解候補率}) = \frac{\sum_{\ell=1}^{8} \text{正答ジェスチャ数}\ell}{\text{検出された全ジェスチャ数}} \tag{2.22}$$

ここで，正答ジェスチャ数ℓは入力画像列ℓ中の20個のジェスチャの内で正しく認識できたジェスチャ数である．また，3フレーム以上連続して同じ認識結果になった場合に「検出」されたとした．

ここで，入力画像サイズ$N_1 = 64$として，特徴ベクトルの次元数$(N_2 \times N_2)$の最適な値を求めるため，$N_2 = \{1, 2, 3, 4, 5, 7, 10, 16\}$と変化させた．ここで得られた$N_2$の最適値を用いて，入力画像サイズ$N_1$を$N_1 = \{3, 6, 9, 12, 15, 30, 64\}$と変化させ最適値を求めた．ここで，衣服および背景の影響を調べるため，

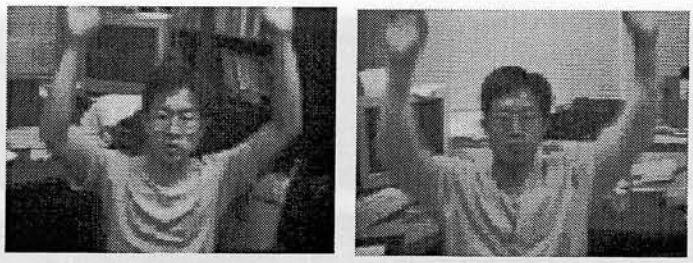
S1 標準パターンの作成時と衣服及び背景が等しい場合

S2 標準パターンの作成時と衣服及び背景の明るさがともに異なる場合

を設定した(図2.10)．衣服の色は，S1のときに灰色，S2のときに黄色であった．標準パターンはS1の場合に作成し，しきい値h_ℓはS1の場合の1位認識率が極力大きくなるよう人手で設定した．S2にはこのS1で作成した標準パターンとしきい値を用いて認識実験を行なった．

2.4.2 実験結果

認識実験の結果を図2.11に示す．衣服と背景が異なる場合(S2)でも，$N_2 = 3, 4, 5$で約80%と高い1位認識率が得られたため，本手法が衣服と背景の変化にロバストであることが示せた．約20%の誤認識の原因は，衣服と背景が異なる場合に生じる，(1)服のしわのでき方の違い，(2)



S1:標準パターンと同じ, S2:標準パターンと異なる

図 2.10 衣服と背景

手の影の違い，(3)着膨れによる人物の大きさの違いが考えられる．計算量を考慮すると，N_2が3のときに今回用いた8種類のジェスチャに対する最適な認識システムとなる．また，N_2が7以上で1位認識率が低下しているが，これは画像の分割数が大き過ぎて動作軌跡の変動を吸収できなかったためと考えられる．

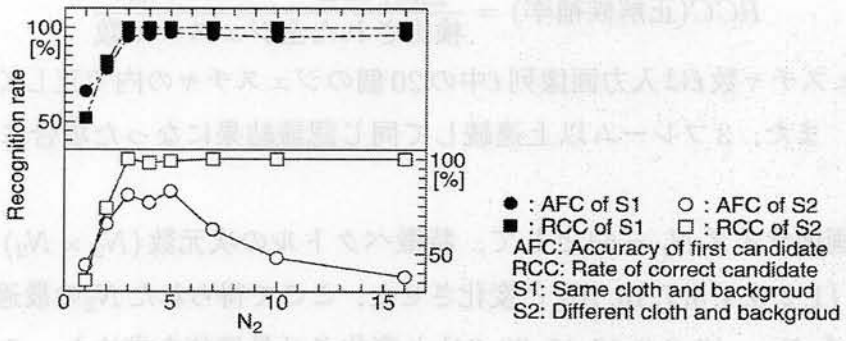


図 2.11 分割数N_2と認識率（$N_1 = 64$）

次に，$N_2 = 3$に固定して，N_1を変化させたときの認識結果を表2.2に示す．この結果から$12 \leq N_1$において約80%の認識率があり，$N_1 \leq 9$では認識率が低下している．$N_2 = 12$のとき，特徴ベクトル値は，$N_1/N_2 = 12/3 = 4$であるため$4 \times 4 = 16$段階となる．このとき認識率が悪化しなかったことから，連続DPによる認識には16段階程度の特徴ベクトル値で十分であると考えられる．従って，$N_1 = 64, N_2 \leq 16$でも特徴ベクトル値は16段階以上になるため，図2.11の実験結果に対する量子化誤差の影響は無視できると考えられる．

以上の結果から，$N_1 = 12$程度と小さな人物画像からでも高い認識率でジェスチャを認識できることが示せた．これは，特徴ベクトルの次元数が3×3（$N_2 = 3$）と小さくても，高い認識率が得られるためである．

表 2.2 入力画像サイズ N_1 と認識率

N_1	3	6	9	12	15	30	64
AFC of S2(%)	50	62	64	80	77	78	81

2.4.3 全方位視野を用いた認識

前節で示したように，本章で提案した動き特徴を用いれば，入力画像サイズが小さくても人物のジェスチャを認識できる．従って，視野を大きく取り低解像度の複数の人物のジェスチャを同時に認識することも可能となるはずである．複数の人物のジェスチャを同時に認識する場合，装置構成を単純化するために，山澤ら [70] が提案した全方位視覚センサ HyperOmni Vision 1台を用いることとする．これは，カメラを上向きに設置して双曲面ミラーを撮影する構成になっており，下方視野を含んだ全方位画像を実時間で得られるだけでなく，画像の補正も容易であるという特長を持つ．例えば，図 2.12(a) では，円卓での会議中に中央に置かれた HyperOmni Vision がすべての出席者のジェスチャを 1 つの画像中で同時に捉えている．また，図 2.12(b) では，自律走行ロボット Nomad の上部に設置された HyperOmni Vision がロボット周辺の複数の人物をとらえている．このように，本節では，人物が移動ロボットへジェスチャで指示を送る場合や計算機が会議中の人物のジェスチャを認識することを想定した実験を行う．

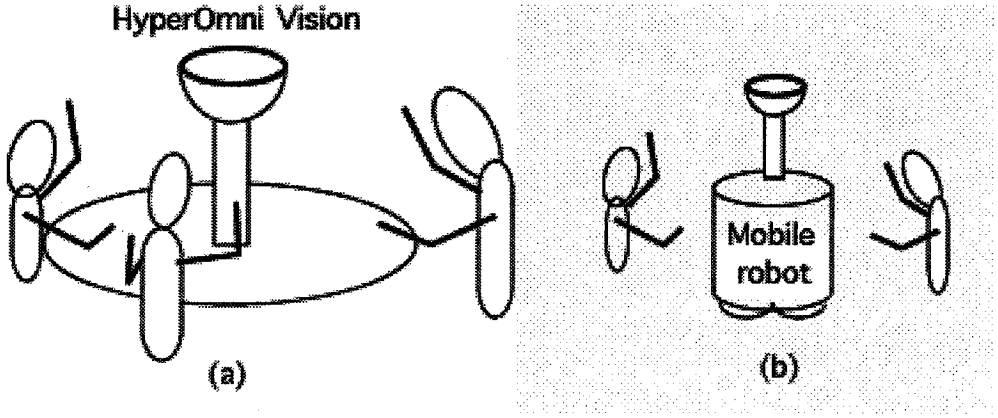


図 2.12 HyperOmni Vision による複数の人物のジェスチャ認識

実験は，自律移動ロボット Nomad 上に HyperOmni Vision[70] を設置し，Nomad の周辺に椅子に座った 4 人の被験者を配置した．各被験者は，Nomad の方向を向いてジェスチャを行う．光源，原画像のサイズ，しきい値などの実験条件は，前節の実験と同様とした．

HyperOmni Vision による 4 人の人物を捉えた映像の一例を図 2.13に示す．各人物の切り出しは人手で行った．さらに，ここで切り出された人物範囲内を 3×3 に等分割し，各分割領域内に重心がある画素を用いて特徴抽出を行った．この図 2.13のように，HyperOmni Vision ま

での距離により人物の大きさが異なるため，各人物の特徴抽出部への入力サイズ$N_1 \times N_1$は異なる．最も離れた人物1までの距離が約4 mであり，そのときの人物の画像サイズは18×15であった．この入力画像から今回提案した特徴抽出法により，3×3次元の特徴ベクトルを算出した．なお，HyperOmni Visionによる歪みは修正していない．

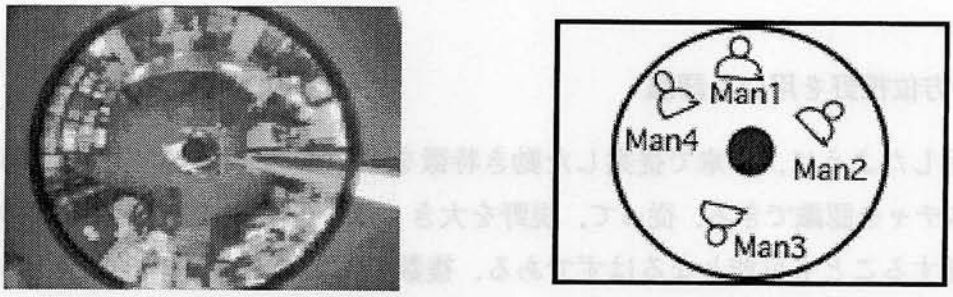


図 2.13 4人の人物動作を捉えたHyperOmni Visionの画像(160×120)および人物の位置（右図）

実験に用いたジェスチャは，前節の実験と同様で8種類とした．標準パターンは各人物について作成した．図2.14に，ジェスチャ“バンザイ”の3フレーム毎の画像系列を示す．入力画像系列は，標準パターンの撮影時と同じ服装にて4人の人物が思い思いにジェスチャを行い撮影した．この入力画像系列のフレーム数は457，この間に4人が行ったジェスチャは10回から13回であった．

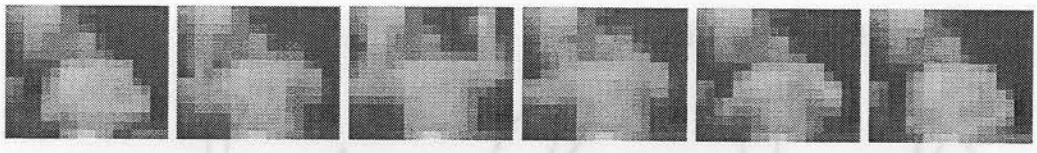


図 2.14 人物1のジェスチャ“バンザイ”(3フレーム毎,18×15)

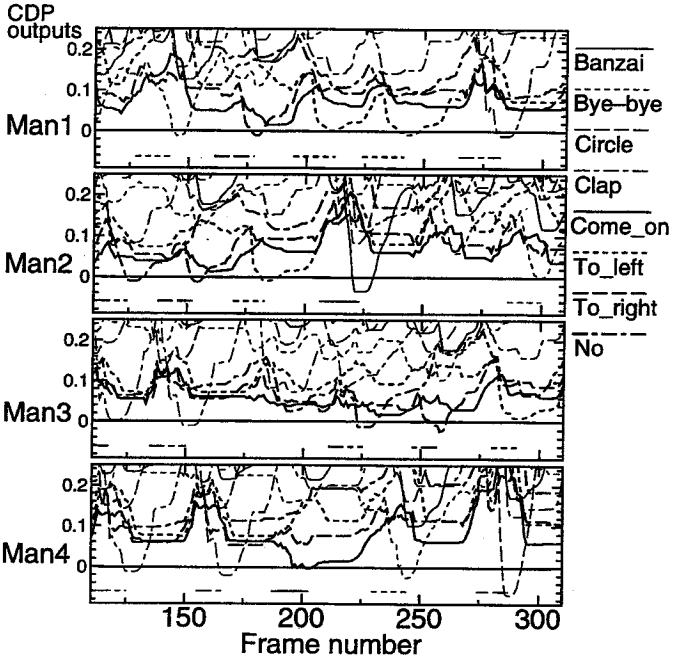
2.4.4 実験結果

表2.3に各人物のジェスチャの認識率を示した．服装と背景が標準パターン作成時と同様であるものの，約80%という高い認識率で認識できており，本手法の有効性が示せた．

表 2.3 各人物のジェスチャの認識率

	Man1	Man2	Man3	Man4
AFC(%)	83	88	82	83

さらに，図2.15に，4人の人物動作に対する連続DPの出力値の例を示す．横軸はフレーム数であり，この上に描かれた横線は，実際に被験者が行ったジェスチャとその時間区間を示している．また，縦軸のCDP出力はそれぞれのしきい値を引いた値である．従って，CDP出力の値が負になった場合に認識されたことになる．この図2.15から，認識もれの場合でも適切な標準パターンのCDP出力が減少していることが分かる．



横軸上の横線：実際に被験者が行ったジェスチャとその時間区間

図 2.15　4人の人物動作に対する連続DPの出力値の例

2.4.5　実時間ジェスチャ認識システム

Indy(R4000 100MHz)を1台を用い，動き特徴のみを用いた実時間ジェスチャ認識システムを作成した(図2.16)．$N_2 = 3$，サンプリングレートは15Hzであり，入力画像を実時間で表示，複数の人物のジェスチャを認識し結果を表示する．この認識システムは，1つの特徴抽出プロセスと各人物に対応した人数個の認識プロセスからなり，この人数は容易に変更可能である．図2.16では，2人の人物を認識対象とした本システムのモニター画面を示した．図2.16の上方の特徴抽出プロセスは，各人物ごとの特徴を抽出し，対応する認識プロセス（図2.16の下方）にソケット通信で送る．各認識プロセスは予め作成したモデルと連続DPでマッチングを行ない，認識結果を表示する．人物の切り出しは，動作中に人物が移動しないことを仮定して，人手で事前に行なった．図2.16中の白枠で表示した範囲を3×3に分割している．計算量としては，Indy 1台で5人程度の人物の動作を認識できる．

図2.17のように，小型全方位ミラーアタッチメント[1]（図2.17の右下）に二人の人物が向かい実時間認識実験を行なった．ジェスチャの種類は，すでに提案した複数の人物間での音声とジェスチャ認識の統合システム [62] を想定し，“左へ”,“右へ”,“前へ”,“後ろへ”,“不要です”, “大きく”,“小さく” の７種類とした．また，モデルは，１回のジェスチャ動画像から得られた特徴ベクトル列を用い，二人のジェスチャモデルは同一のものを用いた．この認識実験の結果，衣服が異なっても約８割の認識率が得られた．

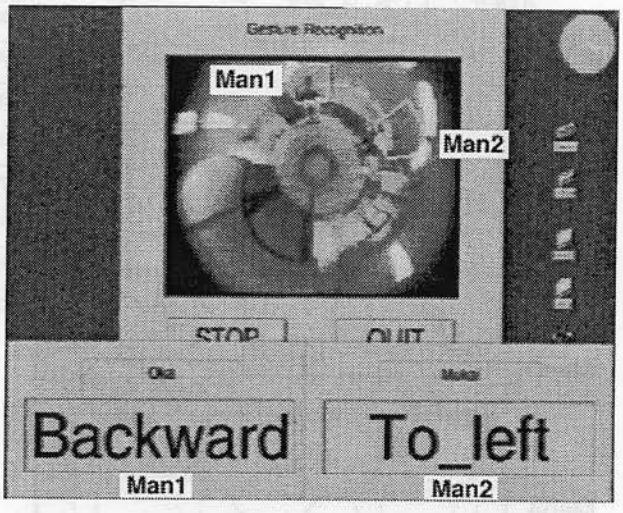


図 2.16 実時間ジェスチャ認識システム (入力画像は上方の白枠)

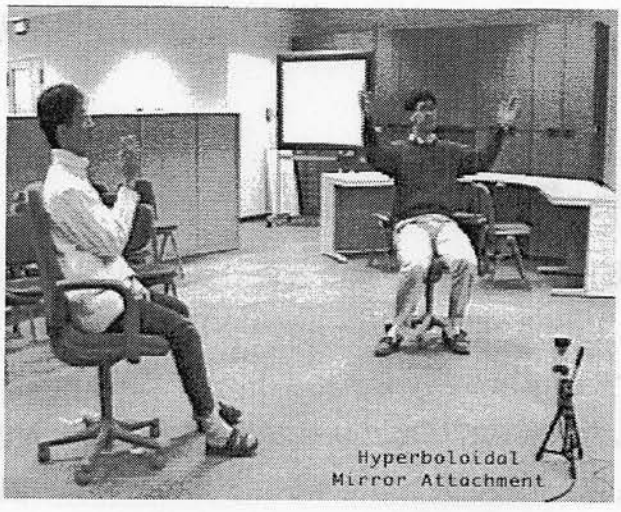


図 2.17 実時間認識実験風景

[1]これは石黒浩氏（京都大学大学院工学研究科情報工学専攻）が，先に双曲面鏡を有する全方位視覚センサを開発した山澤氏（現奈良先端，元大阪大学）[70] らのコメントをヒントに，改良，再設計したもの（詳しくはHome Page: http://www.lab7.kuis.kyoto-u.ac.jp/参照）である．

2.5 動きと形状特徴特徴の評価

2.5.1 実験方法

評価実験で用いた42種類のジェスチャを図2.18に示す．このジェスチャの構成は，“Circle”, “Clap”, “No”, ···, “Inai-ba”, “Batsu” の14種類の動作を基本とし，それぞれの基本動作の右側に示したように手の形を変化させた2種類の動作を加えた．例えば，動作“Circle” の手を縦にして開いた（指が5本）動作を “CiTate5”，動作 “By” の手の形を “グー” で行なったものを “ByGu”，動作 “Come on” の閉じた手（1本の棒状）を縦にした動作を “CoTate1” などと表記してある．これらの中には，“GoYoko5” のように手を横にして相手に示すような指文字に似た動作も含まれている．また，手を波の様に左から右に動かす “Wave” のように，手の方向が時間的に変化するものも含まれている．

また，SGI社のIndy(R4400 200MHz) 1台と附属のIndyComを用いて実時間ジェスチャ認識システム作成した．図2.19の左方は特徴抽出プロセス，図2.19右方は随時教示可能な42単語の認識プロセスのモニタ表示画面である．特徴抽出プロセスは，15フレーム/秒でサイズ160 × 120pixelの画像を取得し動き特徴と形状特徴を抽出して認識プロセスにソケット通信で送る．このシステムの計算量は，1〜2秒のジェスチャ42単語の認識時でもIndyのCPUパワーの約85%であった．

評価実験のために，標準パターンとして42種類のジェスチャを1回ずつ行なったものと，入力パターンとして42種類のジェスチャを10回ずつ行なったものとをビデオに収録した．次に，標準パターン用のビデオを再生して認識システムに教示し，直後に42種類×10回のジェスチャを入力して認識率を求めた．認識率は，正しいジェスチャのみがスポッティングされた数を全入力ジェスチャ数(420)で割って求めた．この実験を，式(2.6)の形状特徴に対する重みw_sを$w_s = 0, 2, 5, 10, 20, 30, 40, 60, 100$と変化させて行ない，認識率を調べた．また，重み$w_{\mathrm{len}}(l)$と$w_{\mathrm{pos}}(i, j, t)$の効果を調べるため，それぞれの重みを用いない場合とどちらも用いない場合についても実験した．

2.5.2 実験結果

認識実験の結果を図2.20に示す．初めに本システムの実験結果について述べる．$w_s = 20, 30$のときに認識率が約80%となり，本章で提案した形状特徴の有効性が示せた．なお，$w_s = 30$の動き特徴と形状特徴のパワーは，ほぼ同じであった．また，“Wave” のように時間的に手形状が変化するジェスチャにも対応できることが示せた．$w_s = 0$のときは，従来の動き特徴のみを用いたことになるが，認識率は約40%に低下した．これは，手の形状が判別できなかったたことが主な原因であった．また，$w_s > 40$で認識率が低下する原因は，動き特徴の寄与率が低下して動きの判別ができなくなったためである．

次に，重み$w_{\mathrm{len}}(l)$と$w_{\mathrm{pos}}(i, j, t)$の効果について考察する．$w_{\mathrm{len}}(l)$を用いないときは，“Circle

図 2.18 評価実験で用いた４２種類のジェスチャ

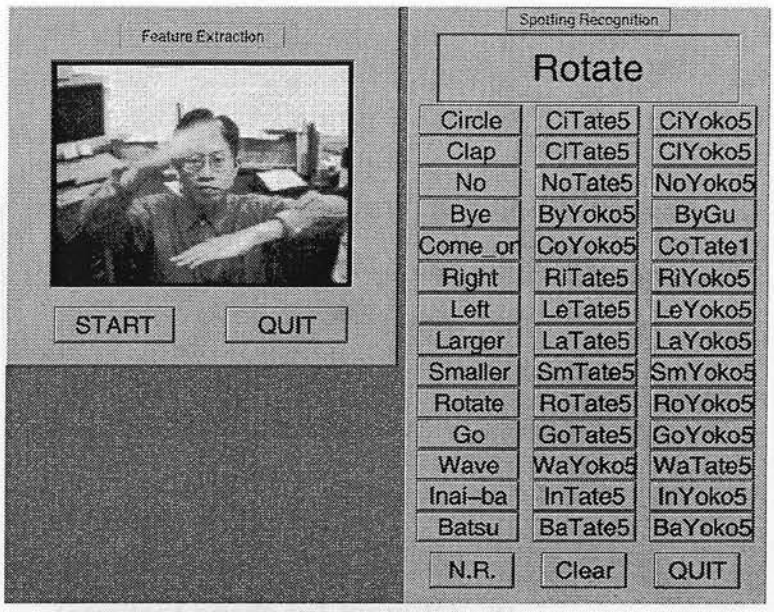


図 2.19 教示可能な実時間ジェスチャ認識システムのモニタ画面

" や "Wave" などの動きの大きなジェスチャの認識率が低下していた．これは，ジェスチャを行うときに体部も移動することが多く，これによって比較的大きな領域が$w_{\mathrm{len}}(l)$で低減されることなく形状特徴に影響を与えたためと考えられる．$w_{\mathrm{pos}}(i,j,t)$を用いない場合には，"Larger" などのように手が画像中の上の方を動くジェスチャの場合に誤認識が多かった．また，$w_{\mathrm{len}}(l)$を用いない場合より若干認識率が低下している．これは，手部以外による領域が画像の下の方に発生しやすく，これらが$w_{\mathrm{pos}}(i,j,t)$によって効果的に低減されたためと考えられる．どちらの重みも用いない場合は，約50%までしか認識率が上がらなかった．

図2.21に各々のジェスチャの認識率を示す．横軸は，図2.18の各ジェスチャのスナップショットの下に示したジェスチャ番号である．"Clap" と "Smaller" では動きが類似しているが，特に(5)ClTate5と(26)SmTate5，(6)ClYoko5と(27)SmYoko5とは手形状も類似しており，これらのジェスチャ間で誤認識が発生していた．また，動きが同じジェスチャの中では，手先の方向が同じで手が開いているものと閉じているものとを混同することがあった．これは，時間差分2値画像の雑音によって輪郭方向の分布に影響が出たためと考えられる．従って，例えば輪郭の平滑化を行った後に方向分布を求めることで形状特徴の定量的な安定性を向上させ，更なる認識率向上が可能であると考えられる．

図2.22には，時間的に手形状（方向）が変化するジェスチャ"Wave" , "WaYoko5" , "WaTate5" を連続して行なったときの形状特徴の時間変化の様子を示した．横軸が画像フレーム番号，縦軸が形状特徴の値であり，図2.22(a)には$f_s^l(k,t)$を図2.22(b)には$f_s^r(k,t)$を示した．この図から，3種類のジェスチャの形状特徴が異なっていることが分かる．

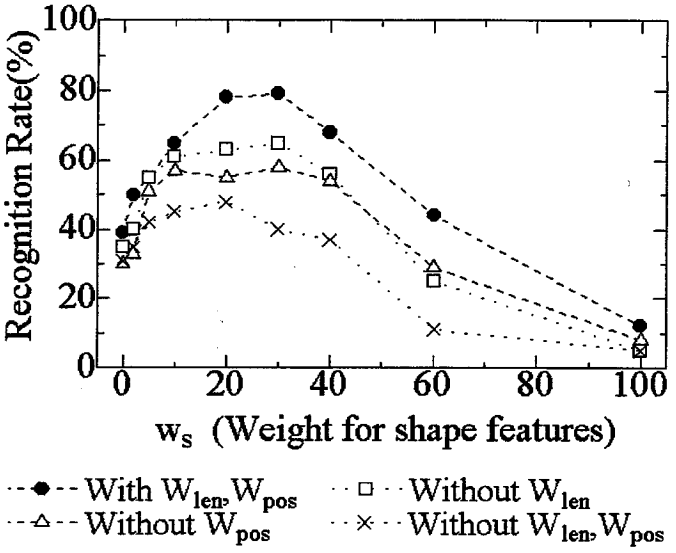
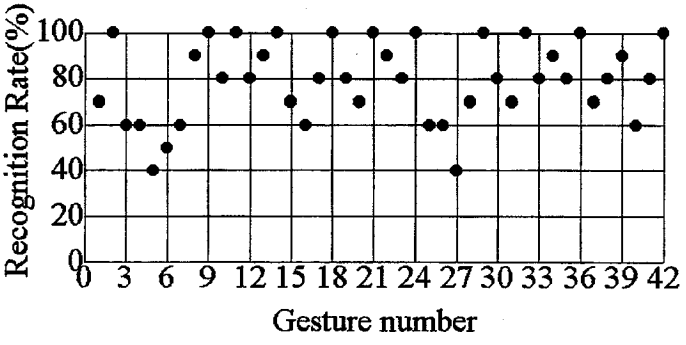
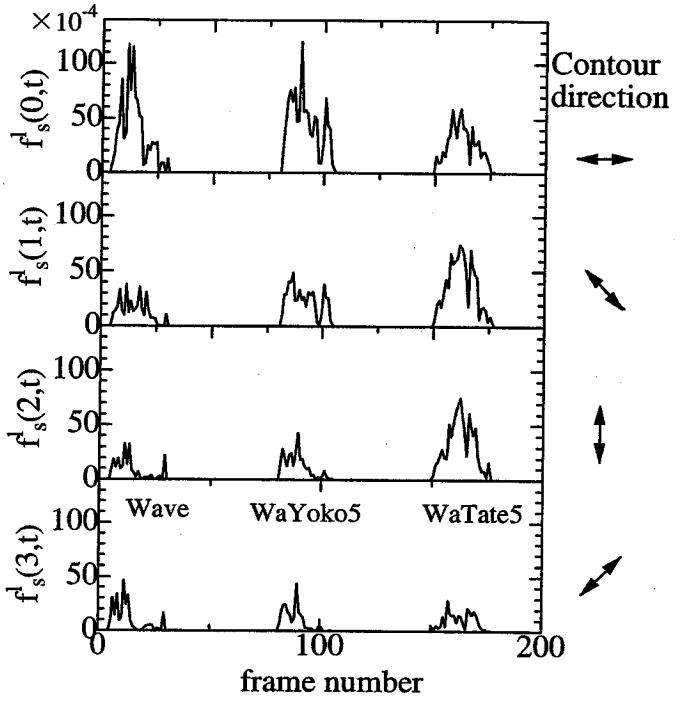


図 2.20 認識実験結果

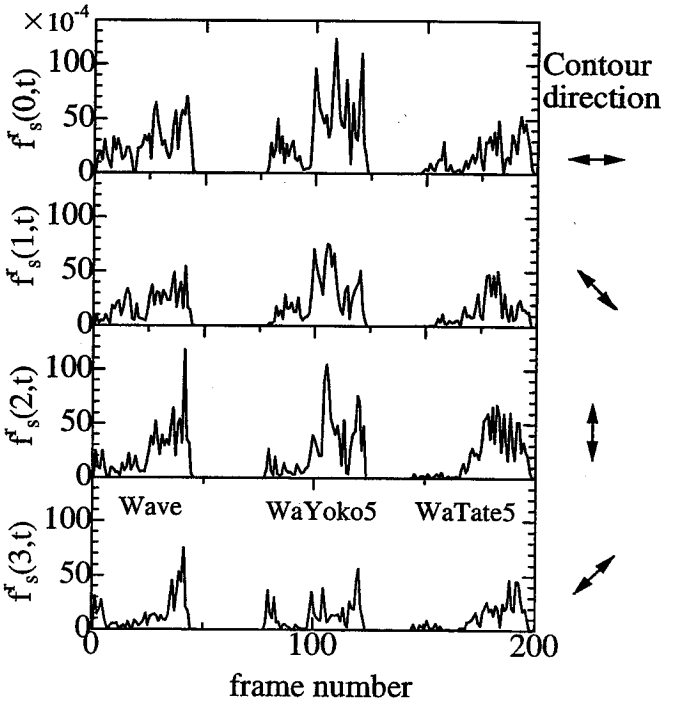


横軸は，図2.16中に示したジェスチャ番号.

図 2.21 各ジェスチャの認識率



(a)



(b)

図 2.22 形状特徴の時間変化の様子

2.6 まとめ

スポッティング認識可能な連続DPを用いたジェスチャ認識システムにおいて，白黒動画像から見かけベースな特徴を抽出する動きと形状特徴を新たに提案した．

動き特徴のみを用いた実験では，8種類のジェスチャについて，画像中の人物サイズが12×12pixel程度でも約8割という高い認識率を実現できた．また，ワークステーション1台で実時間ジェスチャ認識システムを作成し，低解像度で複数の人物のジェスチャ認識を実現した．

動きと形状特徴を用いた実験では，手の形状の判別が必要となる42種類のジェスチャについて，約80%の認識率を達成した．

今後の課題は，画像中の人物領域の切り出しを自動化した上で，ジェスチャの認識率を調べることである．また，形状特徴における課題としては，2次の局所自己相関[6]を導入して形状特徴の高精度化を行なうことが考えられる．

第 3 章

ジェスチャのオンライン教示

3.1 はじめに

同じ意味の動作でも，人物によって大きく変化することが考えられる．また，同一人物でも，その時の体調や気分によって動作が変化することもあろう．この場合，一般的に画像から得られる特徴も変化するため，この変化を特徴抽出法で吸収するのは困難であり，認識手法における考慮が必要となる．もし，動作がさまざまに変化した動画像データが事前に得られれば，認識部において動作変化に対応できるだろう．しかし，一般に事前のデータ収集には時間がかかるか実用上困難であることが多い．このような場合，実時間ジェスチャ認識中にジェスチャ教示を1回行うことで認識が可能となるオンライン教示システムを実現すれば，容易に動作の変化に適応できる．

1.2節で述べたように認識手法であるHMM [27]では，モデル作成のために複数のデータを必要とし，1回の教示直後から認識可能なシステムを作成することが困難である．一方，連続DP[15]では，1個のジェスチャ動画像データからモデルを作成できるため，オンライン教示システムを実現できる．

しかし，高橋らが提案したジェスチャ認識システム [29]では，事前に作成したモデルを固定して用いていたため，オンラインでモデルを変更することは困難であった．もし，モデルを変更したい場合は，システムを停止させ，モデルを含む画像系列を作成し，この中から動作区間を切り出す作業を人手で行うことになる．また，認識基準となるしきい値は，ジェスチャを複数回行い，その認識率が高くなるように人手で調整していた．これは，複数回のジェスチャを行って初めてしきい値を決定できるということである．同様に，連続DPを用いた音声認識においてもオンラインでモデルを教示することは困難であった．

そこで，本章では，モデルの自動切り出し法を提案し，しきい値を実験的に決定することで，オンライン教示可能なシステムを作成して動作者に適応可能とする．また，特徴抽出法の一例として2章で述べた動き特徴を用いて本システムの有効性を示す．さらに，複数の教示データの平均と分散を標準パターンとして用いることで，モデルを教示するたびに認識率が向上する認識システムを構築する．

本章では，3.2節にてオンライン教示システム実現に必要なモデルの切り出し法を提案し，3.3節にて人手を不要とするためにしきい値を決定する．また，3.4節では，オンライン教示可能な実時間認識システムの評価を行なう．3.5節では，認識率向上を目指してモデル平均化法を提案し，3.6節にて評価し，最後に3.7節でまとめる．

3.2 モデルの切り出し

3.2.1 条件設定

モデル切り出しを自動的に行なうために，動作者がモデルを教示するときの手順を以下のように設定する．

「教示手順」：教示したいジェスチャの前後で時間t_s以上静止し，さらに，システムに対して教示の指示を行なうこと．

オンライン教示システムの方では，スポッティング認識を行ないつつ，常時，過去の一定期間の特徴ベクトル列を記憶しておき，教示の指示があったときのみ，モデル切り出しを行なってモデルを更新するものとする．

3.2.2 モデル切り出し

3.2.1節で述べた教示手順に従って教示の指示があった場合には，特徴ベクトル列のパワー$P(t)$

$$P(t) = \frac{1}{N}\sum_{k=1}^{N} u_t(k)^2 \tag{3.1}$$

からモデル部分の切り出しが可能となる．この$P(t)$の変化は，図3.1のようになる．横軸が時間，t_0は教示を指示した時刻である．この図3.1のように，ジェスチャを行なっている間は，途中で動作が短時間静止することがあるもののパワーが高くなる．また，ジェスチャ終了後の静止している間はパワーが減少し，システムに教示を指示しようとして体を動かすとパワーが上昇する．

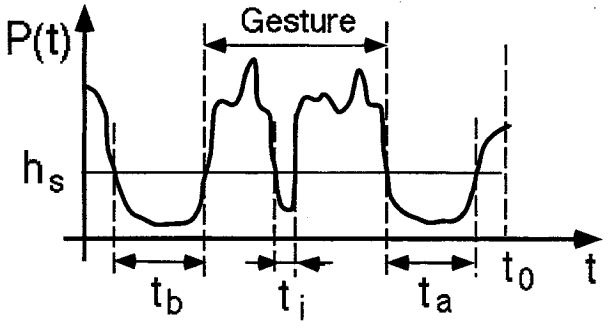


図 3.1 特徴ベクトルのパワー変化を利用したモデル切り出し

従って，パワー$P(t)$の変化は，図3.1のようになり，ジェスチャを行なう前，行なっている途中，行なった後の静止時間をそれぞれ，t_b,t_i,t_aとする．動作者が教示手順に従っていれば，ジェスチャ前後の静止時間t_b, t_aは教示用静止時間t_sに対して，$t_s < t_b, t_a$ を満たすはずである．このとき，登録指示時刻t_0から過去にさかのぼり，しきい値h_s以下である時間t_a以上の区間と時間t_b以上の区間とを検出し，この二つの区間ではさまれる区間をモデルとして切り出せばよい．ただし，$t_i < t_b, t_a$を満たすことが必要であり，もし満たさない場合はジェスチャの一部のみ，または前後の動きを含んだ部分を切り出すことになる．

3.2.3 パラメータ決定のための実験

本節では，実際に動作者に教示を行なってもらい，モデル切り出し法のパラメータ（教示用静止時間t_s，モデル切り出し用しきい値h_s）を決定する．実験装置として，SGI社のIndy(R4400 200MHz)と付属のカメラIndyComを用いた．実験は，オフィス内で椅子に座った1人の被験者に対して行った．カメラの視野は被験者のジェスチャが適切に入るように設定した．また，照明は建物の天井に設置されている蛍光灯のみを用いた．CCDカメラの出力映像をAD変換して得られる画像は，サイズ120×160 ($N_1^i = 120, N_1^j = 160$)，1画素256階調のRGB画像であるが，認識には比較的輝度に強い影響を与えるグリーン成分のみを用いた．画像は15フレーム／秒でサンプリングし，式(2.1)のしきい値 h_cはカメラの熱雑音を考慮し10とした．また，本実験では式(2.2)で求めた特徴を用いた．

実験に用いたジェスチャは，仮想物体の操作や移動ロボットへの指示を想定し，(1)Larger, (2)Smaller, (3)Forward, (4)OK, (5)No, (6)Backward, (7)Left, (8)Right, (9)Bye, (10)Rotate, (11)Go, (12)Stop の12種類とした．図3.2に各ジェスチャのスナップショットを示した．

実験では，3人の動作者が，12種類のジェスチャの教示を行なう動作をしてビデオに収録したものを用いた．ただし，各動作者には，ジェスチャ前後に約1秒以上静止するよう（$t_s = 1$秒）指示した．3人の動作者が行なったジェスチャ“OK” の画像系列を図3.3に示す．ジェスチャのスピードや両手の軌跡が異なっていることが分かる．動作の速度は，動作者3が最も速い．また，動作者1は絶えず両手を接触させているのに対し，動作者3は手を挙げるときに離れている．また，動作者2では手を挙げるときも下げるときも両手が離れている．各動作者は，各ジェスチャを自然な速さで2回，約1.5倍遅くして2回，約2倍遅くして2回の合計6回行った．これは，低速のジェスチャではパワーが減少し，切り出し結果に影響すると考えられるためである．連続DPでは，$\frac{1}{2}$〜2倍の伸縮を入力に許しており，伸縮が$\frac{1}{2}$倍未満か2倍より大きい場合は，原理的に異なったカテゴリーであると判定する．このために約2倍まで遅くしたジェスチャを調べたが，2倍遅い動作は不自然に見えた．

次に，このビデオ映像から特徴ベクトル列を求め，教示の指示を行なった時刻から80フレーム前までのデータを用意した．このデータは，教示したジェスチャの数（12個×6回×3人分）だけ有る．さらに，これらの特徴ベクトル列のパワーを求めた．

図 3.2 12種類のジェスチャのスナップショット

(a) 動作者 1 ,(b) 動作者 2 ,(c) 動作者 3 .

図 3.3 人物による動作の違い．ジェスチャ“OK”(5フレーム毎) の場合.

しきい値h_sの最適値を決定するために，しきい値を0.001, 0.003, 0.01, 0.03, 0.01と変化させて区間検出率を求めた．この区間検出率は，正しい切り出し区間をT，しきい値h_sによって切り出される区間をRとして，

$$\text{区間検出率} = \frac{T \cap R}{T \cup R} \tag{3.2}$$

で求めた．ただし，t_b, t_aは0.6秒とした．

3.2.4 実験結果

図3.4に，しきい値h_sを変化させたときの区間検出率を求めた．通常のジェスチャの場合では，しきい値が0.003〜0.01のときに区間検出率が96%と最も高くなっている．また，ジェスチャの速度が遅い場合には，パワーが減少し高いしきい値のときに区間検出率が減少した．しかし，図3.4に示すように2倍遅いジェスチャを行った場合でも しきい値が0.003のときに区間検出率が94%と高かった．そこで，今後は，$h_s = 0.003$として実験を進める．一方，このときの動作途中の静止時間t_iの最大は，約0.7秒（10フレーム），ジェスチャ前後の静止時間t_b, t_aは最小でも0.8秒（12フレーム）であった．従って，適切に切り出し可能な必要条件$t_i < t_b, t_a$を満たしている．今後，教示用静止時間t_sは，若干の余裕をみて約1秒とする．

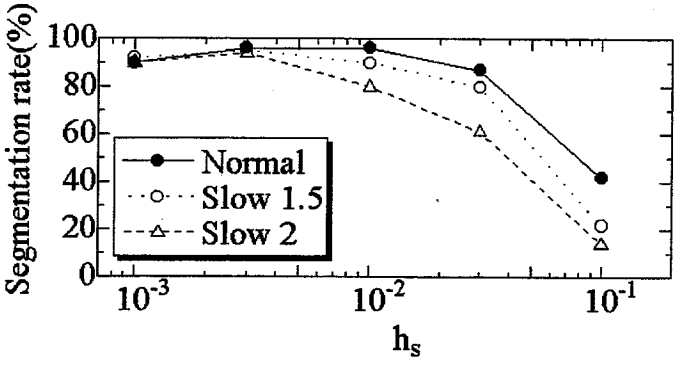
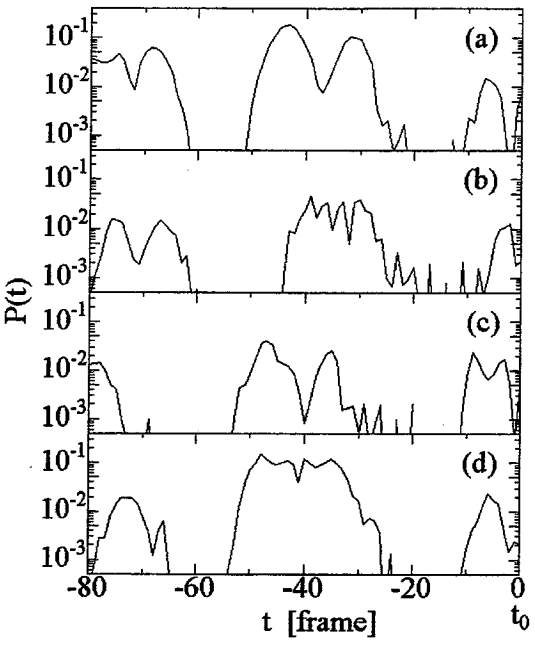


図 3.4 しきい値h_sを変化させたときの区間検出率

図3.5には，一例として動作者1が自然な速さで行なった4個のジェスチャ教示におけるパワーの変化の様子を示した．図3.5(a)は，ジェスチャ“Larger”の教示を行なったものであり，短時間であるが動作の途中で動きが止まっている．このことは，ジェスチャ“Smaller”，“OK”，“No”，“Go”，“Stop” でも同様であった．図3.5(b)は，ジェスチャ“Forward”によるものであるが，動作の途中で小刻みに静止するためにパワーが波だっている．ジェスチャ“Backward”，“Bye”でも同様な変化が見られた．図3.5(c)は，ジェスチャ“Left”によるものであるが，ジェスチャ“Right”でも同様であった．図3.5(d)は，手を胸の位置で回転させるジェスチャ“Rotate”であり，動作中は絶えず大きなパワーとなっている．



(a)"Larger",(b)"Forward",(c)"Left",(d)"Rotate".

図 3.5 自然な動作で教示を行なったときの特徴ベクトルのパワーの変化

3.3 しきい値の決定

3.3.1 正規化距離の導入

同一人物が同じジェスチャを行なっても，椅子の位置や体調によって若干変動する．しかし，式（2.20）のしきい値h_ℓを，この変動量の最大値より大きく設定すれば，ジェスチャの検出洩れが無くなる．従って，各ジェスチャの変動量が予め分かっていれば，しきい値h_ℓを決定できることになる．ところが，本システムでは1回の教示直後から，そのジェスチャを認識できるようにしたいため，そのような変動量は得られない．

しかし，複数のジェスチャの変動量を調べた結果，この変動量は該当ジェスチャのパワーにほぼ比例することが分かった．つまり，$\ell(1 \leq \ell \leq L)$番めのモデルのフレーム長をT_ℓ，変動量の最大値をσ_ℓ，パワーをP_ℓとすると，

$$\sigma_\ell \fallingdotseq \alpha \cdot P_\ell \tag{3.3}$$

と書ける．ただし，P_ℓは

$$P_\ell = \frac{1}{T_\ell} \sum_{\tau=1}^{T_\ell} \frac{1}{N} \sum_{k=1}^{N} z_\tau(k)^2 \tag{3.4}$$

と求める．さらに，このσ_ℓは分散値に比例するとも考えられるため，式（2.20）の代わりに以

下の式で認識結果を判定する方が適切である．

$$\ell^*(t) = \begin{cases} \mathrm{Arg}[\min_\ell\{A_\ell(t)/\sigma_\ell\}] & \text{if } \exists\ell \quad \text{so that} \quad A_\ell(t) \le \sigma_\ell \\ \text{null} & \text{otherwise} \end{cases} \tag{3.5}$$

ここでは，モデルとの距離$A_\ell(t)$をσ_ℓで割って正規化した値（正規化出力）を用いて，他のモデルと比較している．従って，事前に式（3.3）のパラメータαが適切に求められれば，式（3.5）によって自動的に認識判定が行なえることになる．

3.3.2 パラメータαの決定

本節では，ジェスチャ動画像から式（3.3）のパラメータαを求める．はじめに，3.3.1節で用いた12種類のジェスチャを10回行ない，ジェスチャのパワーP_ℓと最大変動量σ_ℓとの関係を調べた．これは，3人の動作者について2回行なった．ただし，変動量は，連続DPで得られる累積距離の最小値とし，最大変動量は10個のジェスチャの中で最も変動量の大きいものとした．

図3.6に実験結果を示す．横軸がパワーP_ℓ，縦軸が最大変動量σ_ℓである．この図から

$$\sigma_\ell < 0.15 \cdot P_\ell \tag{3.6}$$

であることが分かる．このことから，$\alpha = 0.15$とすれば，すべてのジェスチャの変動より大きなしきい値となる．勿論，図3.6から分かるように，変動量がこのしきい値より かなり小さいものもあり，カテゴリー外の動きを誤検出する可能性が増すことになが，一回の教示直後から，そのジェスチャを認識できるようにするためには避けられないことである．そこで，次節では$\alpha = 0.15$としてオンライン教示システムを作成し，本手法の性能を検証する．

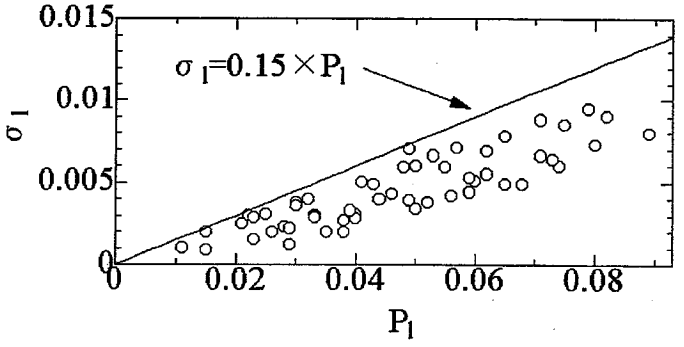


図 3.6 ジェスチャのパワーP_ℓと最大変動量σ_ℓ

3.4 オンライン教示システムの評価

3.4.1 オンライン教示実験

今回実現したシステムのモニター画面を図3.7に示す．左側が特徴抽出プロセスであり，右側の認識プロセスに特徴ベクトルをソケット通信で送っている．この認識プロセスは，上から順に認識結果表示部，登録対象のジェスチャ名ボタン(3.2.3節で用いた12種類のジェスチャ)，終了時のQUITボタンである．教示は，該当するジェスチャ名ボタンをマウスでクリックすることで実行される．計算量は，Indy(R4400 200MHz)のCPUパワーの約65%であった．また，教示では，モデルの切り出しと認識判定用のモデルのパワーを算出することが必要であるが，この計算でシステムが停止する時間は0.1秒以下であり，動作者に不便を感じさせるには至っていない．

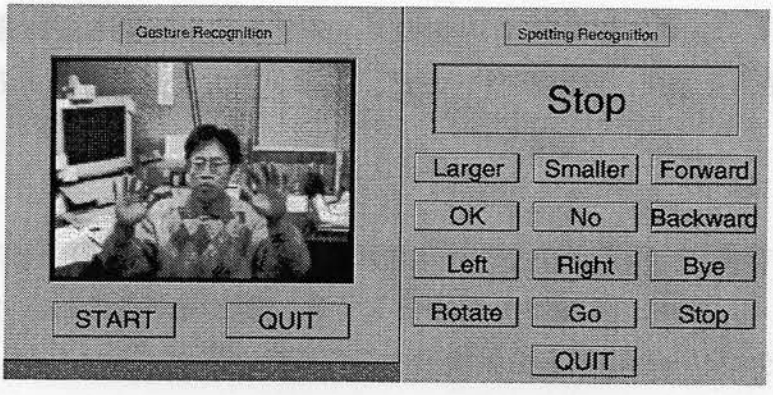


図 3.7 本システムのモニター画面

3.2.3節で用いた12種類のジェスチャを用いて従来システムと本教示システムとを比較した．実験では，3人の動作者が12種類のジェスチャを1回づつ行ってビデオに収録したものをモデル教示用とした．また，3人の動作者が12種類のジェスチャを行なうことを10回繰り返してビデオに収録したものを認識実験用の入力映像とした．従って，入力画像は120個のジェスチャ×3人分となり，合計約15分の録画映像となった．

従来の実時間認識システムでは，事前に動作者1,2,3の教示用ジェスチャ映像からモデルを人手で切り出し，それぞれの動作者の入力用ジェスチャと比較してしきい値を調整した．そして，3人のジェスチャ入力時に以下の3つの場合について認識率を求めた．

[S1]対応する動作者のモデルを切り替えて用いた．

[S2]対応する動作者のモデルを切り替えるが，各モデルのしきい値は本章で提案した認識判定法で設定した．

[S3]動作者1のモデルを固定して用いた．

この中で，[S3]の場合のみが，従来のシステムを停止せずに認識を継続している．

表 3.1 従来システムの認識実験結果
S1:オフライン，モデルは動作者ごとに変更，S2:オフライン，改良した認識判定法でしきい値決定，S3:オンライン，モデルは動作者1で固定．(Man1,2,3と順次実施)

	Man1	Man2	Man3
S1(%)	86	85	84
S2(%)	84	82	85
S3(%)	86	41	52

また，オンライン教示システムでは，動作者1の教示用ジェスチャ映像をオンライン教示後，このシステムに入力映像を入力して動作者毎に認識率を求めた．このとき，正しいジェスチャを認識できなかった場合には教示を行った．認識率は，正答した数を入力ジェスチャ総数120で割って動作者毎に求めた．ここで，入力したジェスチャと一致した認識結果のみを出力した場合を正答とした．

3.4.2 実験結果

従来システムの認識実験結果を表3.1に示した．S1，S2の場合は，動作者に対応したモデルを切り替えて使用しているために各動作者の認識率は高くなっている．これに比べて，S3の場合はシステムを停止することなく動作者1のモデルを固定して使用しているため，動作者2，3の認識率が低くなった．また，本章で提案した認識判定法を用いたS2の場合と従来の人物によるしきい値調整をしたS1の場合とを比較すると大きな差が認められない．これは，本認識判定法の有効性を示している．

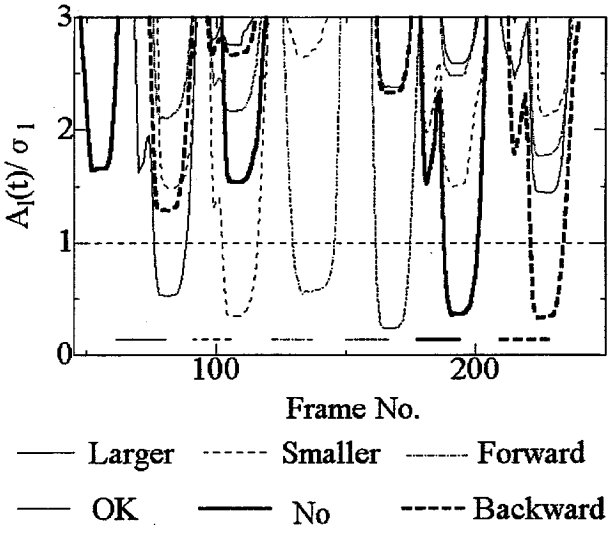
また，オンライン教示システムにおける認識率，区間検出率，教示回数を表3.2に示した．この結果より，動作者が変化しても80%以上の認識率を維持しており本システムの有効性が示せた．区間検出率は，90%以上となっており切り出し法の有効性も示せた．動作者1では教示回数が16回となっているが，原因としては，ジェスチャ“Smaller” と “No” とを混同したりジェスチャ“Forward” を検出できなかったためである．同様な誤検出は従来法でも生じていた．また，動作者2での教示回数が動作者1より多くなっているが，これは動作者2が動作者1のモデルの内7個を更新して動作者2用のモデルを作成したためである．動作者2に関しても一旦教示したジェスチャが認識できず，再教示を11回行っていた．同様に，動作者3でも動作者2のモデルを8個更新し，再教示を12回行っていた．

図3.8には，オンライン教示システムにて6種類のジェスチャを順次行なったときの認識の様子を示した．横軸はフレーム番号であり，この上に描かれた横線は，実際に被験者が行ったジェスチャとその時間区間を示している．また，縦軸は正規化出力 $A_\ell(t)/\sigma_\ell$ であるため，この値が

表 3.2 オンライン教示システムの認識実験結果
上段から認識率，区間検出率，教示回数 (Man1,2,3 と順次実施)

	Man1	Man2	Man3
Recognition rates(%)	89	85	83
Segmentation rates(%)	93	97	94
Number of online teaching	16	18	20

1 以下のときに認識したと判定する．この図3.8から，該当ジェスチャの正規化出力が大きく減少していることが分かる．



正規化出力 $A_\ell(t)/\sigma_\ell$ が 1 以下のときに認識したと判定する．

図 3.8 認識の様子

3.5 モデル平均化法

前節までの提案によって，一回のジェスチャ教示直後から認識可能なオンライン教示システムを実現した．しかし，すべてのモデルが 1 回の教示データから作成されていたため，動作の変動に有効に対処できなかった．認識手法である HMM では，モデル作成のために複数のデータを統計処理して標準的なモデルを作成することで認識率を向上させている．そこで，1 個のジェスチャ動画像データからモデルを作成できるという連続 DP の利点を生かしつつ，平均的な動作をモデルに反映する手法を提案する．

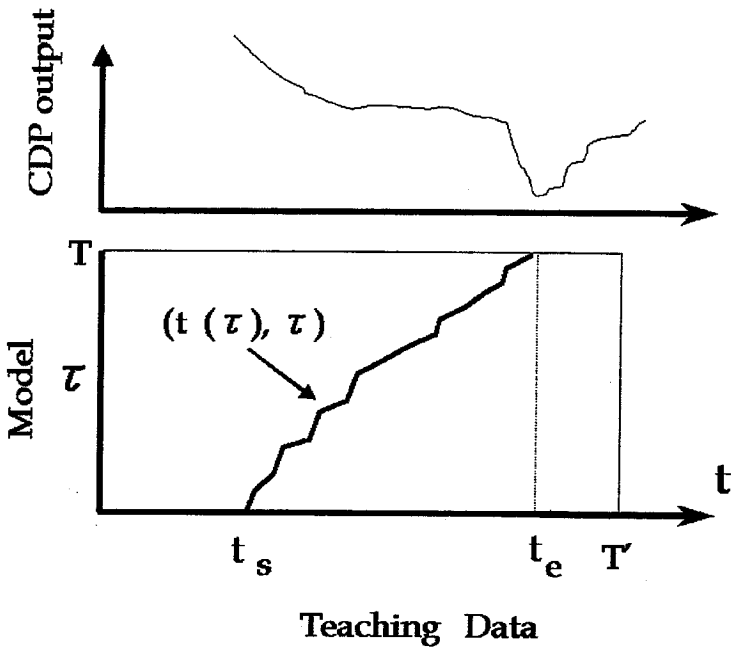


図 3.9 連続DPによって求まった最適パス

3.5.1 モデルの平均化手法の概要

オンラインでモデルの平均化処理を行うためには，動作者は，従来システムと同様に教示動作を行い，該当するモデルを指定し，平均化処理をシステム側に要請するものとする．このとき，システム側では以下の処理を行う．

[1] 該当モデルと新たな教示データとのフレーム対応を連続DPにより求める．このとき，累積距離がしきい値thr_ℓ以上であれば，平均化を行わず，教示データを新たな標準パターンとして登録し，これが認識された場合にも該当モデル名を出力する．

[2] 対応するフレーム毎にモデルの平均化処理を行う．

このように平均化処理を行い，実時間認識には，平均化情報を利用した局所距離算出法を利用する．

3.5.2 連続DPによる対応づけ

本節では，既にM個のデータを平均化して作成したモデルと新たな教示データとのフレーム対応を連続DPにより求める手法について述べる．

通常は，モデル$z_{\tau,k}(1 \leq \tau \leq T, 1 \leq k \leq N)$と入力データとの累積距離を連続DPによって求めているが，本節では，入力データではなく新たな教示データ$u'_{t,k}(1 \leq t \leq T')$との距離を求める．そして，図3.9のように，連続DPの出力値$A(t)$が$(T/2 \leq t \leq T_n)$において極小となる時刻t_eを求める．この時刻が，モデルと最適にマッチングする区間の終点の時刻である．

一方，連続DP算出時に3個のパスを比較しているが，このとき決定したパスの番号$p_{t,\tau}(p_{t,\tau} =$

$\{1,2,3\}, 1 \leq t \leq T', 1 \leq \tau \leq T)$ を記憶しておく．そして，時刻t_eから$p_{t,\tau}$の情報を基にパスを遡れば，最適パス$(t_o(\tau), \tau)(1 \leq \tau \leq T)$ を求めることができ，これが，モデルと教示データとのフレーム毎の対応関係となる．ただし，図2.7のパス1の場合は，τに対応する教示データのフレームが2つとなる．この場合は，大きいフレーム番号を$t_o(\tau)$とする．

このとき，次節で行うモデルの平均と分散を考慮し，モデルのτ番目のフレームに対応する教示データの特徴値$u''_{\tau,k}$ （以後，対応教示データと呼ぶ）を以下のように定義する．

$$u''_{\tau,k} = \begin{cases} \frac{1}{2}\{u'_{t_o(\tau),k} + u'_{t_o(\tau)-1,k}\} & \text{if local path1} \\ u'_{t_o(\tau),k} & \text{otherwise.} \end{cases} \tag{3.7}$$

3.5.3 モデルの平均化処理

M個の対応教示データを$u''_{\tau,k}(m)(1 \leq m \leq M)$，このM個を平均化したモデルを$z_{\tau,k}(M)$，モデルの分散を$\sigma_{\tau,k}(M)$と表記し，以下の式で求めるものとする．

$$z_{\tau,k}(M) = \frac{1}{M}\sum_{m=1}^{M} u''_{\tau,k}(m) \tag{3.8}$$

$$\sigma_{\tau,k}(M) = \frac{1}{M}\sum_{m=1}^{M}\{u''_{\tau,k}(m) - z_{\tau,k}(M)\}^2 = \frac{1}{M}\sum_{m=1}^{M} u''_{\tau,k}(m)^2 - z_{\tau,k}(M)^2 \tag{3.9}$$

実際のシステムでは，過去の教示データ$u''_{\tau,k}(m)$ $(1 \leq m \leq M)$を保持しておく必要は無い．$z_{\tau,k}(M)$，$\sigma_{\tau,k}(M)$ および，新たな対応教示データ$u''_{\tau,k}(M+1)$から，以下の式で$z_{\tau,k}(M+1)$，$\sigma_{\tau,k}(M+1)$ を求める．

$$z_{\tau,k}(M+1) = \frac{1}{M+1}\{M \cdot z_{\tau,k}(M) + u''_{\tau,k}(M+1)\} \tag{3.10}$$

$$\sigma_{\tau,k}(M+1) = \frac{1}{M+1}[M\{\sigma_{\tau,k}(M) + z_{\tau,k}(M)^2\} + u''_{\tau,k}(M+1)^2] - z_{\tau,k}(M+1)^2 \tag{3.11}$$

これらの処理を，$(1 \leq \tau \leq T, 1 \leq k \leq N)$について行い，モデルの平均および分散を求める．

3.5.4 平均化情報を利用した局所距離

モデルの平均および分散が得られているため，以下の式で局所距離$d(t,\tau)$を求める．

$$d(t,\tau) = \frac{1}{N}\sum_{k=1}^{N}\frac{\{u_{t,k} - z_{\tau,k}(M)\}^2}{\sigma'_{\tau,k}(M)}. \tag{3.12}$$

ただし，

$$\sigma'_{\tau,k}(M) = \beta \cdot \sigma_{\tau,k}(M) + (1-\beta)z_{\tau,k}(M). \tag{3.13}$$

とする．これは，教示回数が少ないと分散の値が不正確となることが考えられ，その場合は，近似的に平均値を用いることができるようにするためである．$\sigma'_{\tau,k}(M)$は，$\beta=1$の場合に分散そのものとなり，$\beta=0$の場合に平均となる．さらに，次節の評価実験では$\sigma'_{\tau,k}(M)=1$として，平均情報のみを用いる場合とも比較する．

また，教示データとモデルとの対応付けにおいては，式(3.12)の$u_{t,k}$を$u'_{t,k}$に置き換えるものとする．ただし，対応フレームを求める場合は，$\sigma'_{\tau,k}(M)=1$として平均情報のみを用いた．

3.6 モデル平均化法の評価

3.6.1 実験方法

実験は，3.2.3節と同様な設定とモデルを用いた．教示用のビデオは，12種類のジェスチャを20回繰り返して作成した．このとき，約1秒の時間間隔を置いてジェスチャを行った．また，認識実験用のビデオは，12種類のジェスチャを連続的に10回行って作成した．いずれの場合も自然な範囲ではあるが意図的にジェスチャの変動が生じるようにした．この実験では，正答数を入力ジェスチャ総数(12×10)で割った値を認識率とした．ただし，入力ジェスチャと同じ認識結果のみを出力した場合を正答とした．

実験では，各モデルの教示回数を，1,5,10,20回と変化させて認識率を調べた．このとき，ジェスチャが多少変動しても平均化処理が行われるよう，3.5.1節で述べたしきい値thr_ℓを，$\mathrm{thr}_\ell=2\cdot\sigma_\ell$とした．この条件で教示を行った結果，すべてのジェスチャがこのしきい値thr_ℓ以下の距離となった．

3.6.2 実験結果

図3.10に，従来法と本手法との比較実験の結果を示す．従来法の認識率は，62%と低いのに対して本手法では，変動の大きなジェスチャ12種類のジェスチャ認識にもかかわらず，90%以上となり本手法の有効性を示せた．ただし，今回の実験では，局所距離を分散値で正規化した場合の効果が得られず，平均値のみを用いた方が認識率が高くなっている．この原因としては，今回用いた低解像度特徴の各次元が画像を3×3に分割して得られる情報に起因しており，手の移動に伴って，1つの次元だけを考えた場合の連続性が低いためと考えられる．例えば、手の重心の画像中の位置などは連続的に変化するため，分散値による正規化の効果も生じるものと考える．

図3.11にジェスチャ“larger”　の平均と各教示データを示した．水平の軸は，3番目の特徴ベクトル要素，垂直の軸は6番目の要素である．この図では、5、10、20回の平均と各教示データを示している．同様に，図3.12ジェスチャ“OK”の特徴ベクトルを示した．図3.11，3.12から分かるように平均のパターンが各教示データの平均に収束している．

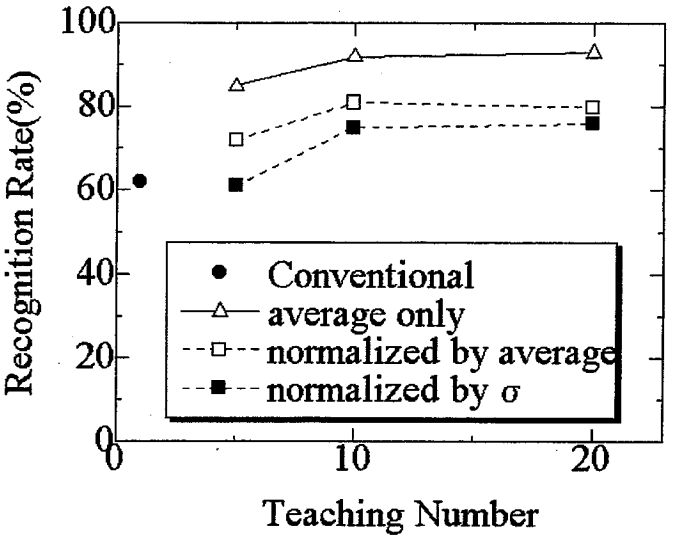


図 3.10 比較実験結果

3.7 まとめ

本章では，モデルの自動切り出し法と実験的なしきい値の決定法を提案し，複数の動作者のジェスチャ動画像をもとに各手法のパラメータを最適化した．また，オンライン教示可能な実時間認識システムを作成して，動作者に適応できることを示した．さらに，複数の教示データの平均と分散を標準パターンとして用いることで，モデルを教示するたびに認識率が向上する認識システムを構築した．

今回の提案手法では，1つのモデルが他のモデルに含まれている場合，どちらのモデルも検出される．このことを防ぐため，5章で述べる重み減衰型RIFCDPによりあらかじめモデル間で類似区間を検出するシステムの実現が今後の課題である．

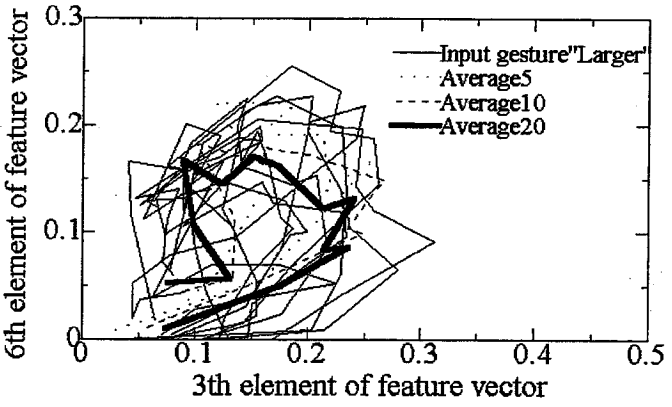


図 3.11 ジェスチャ “larger ” のモデル

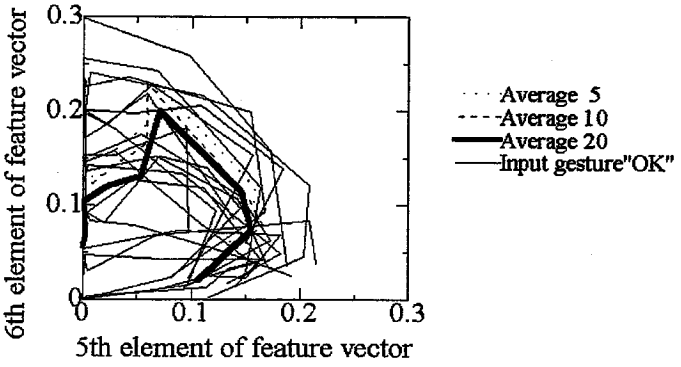


図 3.12 ジェスチャ “OK” のモデル

第 4 章

非単調連続DP

4.1 はじめに

人間のジェスチャは，同一動作であっても途中で戸惑ったり考えて止まったりすることがある．この戸惑っている動作は，時と場合によって無数に変化すると考えられる．従って，もし，このような動作すべてを連続DP[15]で認識しようとすると，多くの標準パターンが必要となり非効率的である．

そもそも連続DPでは，従来から順序の逆転の無い音声やジェスチャの時系列データを扱っているため,「傾斜制限」とよばれる単調増加な局所パスのみが取られている．この意味で，連続DPは,「単調」であるといえる．そして，単調であるがゆえに，無数に変化する戸惑い方を認識するために，多くの標準パターンを必要とする．

一方，特徴空間において，戸惑う動作の軌跡を観察すれば，1つの曲線上を移動していると考えられる．ただし，順方向に移動するだけでなく静止したり逆方向に移動しているだろう．このため，認識に必要な情報は，この1つの曲線のみであると考えられる．従って，連続DPで必要となる複数の標準パターンは，大部分が重複した情報となっているといえる．

そこで，この1つの曲線を標準パターンとし，増加だけでなく減少をも認める局所パスを導入すれば，必要最低限な情報で認識が可能となる．これは，連続DPに非単調性を導入したものといえるため,「非単調連続DP」と呼ぶ．

しかし，従来の連続DPにおいて減少を認める局所パスを導入すると，マッチングしたパスの長さが大幅に変化し，パスの始点から終点までの累積重みを算出できなくなる．これによって，パスの長さを正規化できず，正しい累積距離が得られない．そこで，局所距離にかかる重みを，入力パターンにおいて過去に溯るに従って指数関数的に減少させて累積距離を求めることで，この問題を解決する．

また，本章では，環境を撮影した画像系列からなるトポロジー地図を用いた，移動ロボットの位置推定においても非単調連続DPが有用であること示す．従来は，スポッティング的な位置推定が可能なReference Interval-Free 連続DP(RIFCDP)[60]手法を用いていたが，移動ロボットの移動方向が地図作成時と同じ方向に制限されていた[75]．これは，RIFCDPも連続DPと

同様に「単調」な局所パスしか許していないためである．従って，非単調な局所パスを導入した非単調連続DPを用いれば，ロボットが地図作成時と逆方向に走行したときや停止時でも位置推定が可能となる．

本章では，4.2節で非単調連続DPを提案し，4.3節にて戸惑い動作の認識へ適用する．さらに，4.4節にて移動ロボットの位置推定へ適用し，4.5節でまとめる．

4.2 非単調連続DP

4.2.1 定式化

一つの標準パターンZを，標準動作を捉えたTフレームの動画像から得られる特徴ベクトルz_τの系列

$$Z = \{z_\tau | 1 \leq \tau \leq T\} \tag{4.1}$$

で表す．ここで，特徴ベクトルz_τはその次元数をNとすると

$$z_\tau = (z_\tau(1), z_\tau(2), \cdots, z_\tau(N)) \tag{4.2}$$

である．入力画像からも同様な特徴ベクトル系列$u_t(0 \leq t < \infty)$が連続的に得られる．このとき，u_tとz_τとの局所距離を$d(t,\tau)$と表記する．この$d(t,\tau)$の定義の一例を以下に示す．

$$d(t,\tau) = \frac{1}{N}\sum_{k=1}^{N}(u_t(k) - z_\tau(k))^2. \tag{4.3}$$

ここで，入力，標準パターンの時間軸をそれぞれt,τと区別する．

さらに，点(t,τ)を終点とした標準パターンと入力系列との累積距離を$S(t,\tau)$で表す．非単調連続DPでは$S(t,\tau)$を以下のような漸化式で更新する．

初期条件$(t=0)$：

$$S(0,\tau) = d(0,\tau). \quad (1 \leq \tau \leq T) \tag{4.4}$$

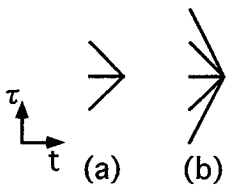
漸化式$(1 \leq t)$:

$$S(t,\tau) = \alpha \cdot d(t,\tau) + (1-\alpha) \cdot \min_{m \in \{-1,0,1\}} S(t-1, \tau+m). \quad (1 \leq \tau \leq T) \tag{4.5}$$

ここで，αは正規化係数$(0 \leq \alpha \leq 1)$である．式を簡単にするために，入力の系列は標準パターンと比べて−1〜1倍の伸縮があってもマッチング可能であるとした．これは，図4.1(a)のような傾斜パターンを採用していることになる．しかし，式(4.5)のmの範囲を変えれば図4.1(b)のような様々な傾斜パターンを設定できる．

ここで 整数$p_0, p_1, \cdots, p_t$ を以下のように定義する．

$$\begin{aligned} p_t = \tau, \quad & |p_k - p_{k-1}| \leq 1 (k = t, t-1, \cdots, 1) \\ & \text{and} \\ 1 \quad & \leq p_k \leq T (k = 0, 1, \cdots, t) \end{aligned} \tag{4.6}$$



(a) $m \in \{-1, 0, 1\}$, (b) $m \in \{-2, -1, 0, 1, 2\}$

図 4.1 非単調連続DPの傾斜パターン例

このとき，式(4.4)，式(4.5)の漸化式は次式のように変形できる．

$$\begin{aligned} S(t,\tau) &= \min_{\{p_{t-1},p_t\}} \{\alpha \cdot d(t,\tau) + \alpha \cdot (1-\alpha) \cdot d(t-1, p_{t-1}) \\ &\quad + (1-\alpha)^2 \cdot \min_{m\in\{-1,0,1\}} S(t-2, p_{t-1}+m)\} \\ &= \min_{\{p_{t-2},p_{t-1},p_t\}} \{\alpha \cdot d(t,\tau) + \alpha \cdot (1-\alpha) \cdot d(t-1, p_{t-1}) \\ &\quad + \alpha \cdot (1-\alpha)^2 \cdot d(t-2, p_{t-2}) + (1-\alpha)^3 \cdot \min_{m\in\{-1,0,1\}} S(t-3, p_{t-2}+m)\} \\ &= \cdots \\ &= \min_{\{p_0,p_1,\cdots,p_t\}} \{\sum_{k=1}^{t} \alpha(1-\alpha)^{t-k} \cdot d(k, p_k) + (1-\alpha)^t \cdot d(0, p_0)\}. \\ &\qquad (1 \le t) \end{aligned} \tag{4.7}$$

つまり，非単調連続DPでは，図4.1(a)の様に(t,τ)において$(t-1,\tau-1),(t-1,\tau),(t-1,\tau+1)$の各点から局所最適パスがとられ，図4.2の実線のように(t,τ)平面での最適パスのτがtに関して単調に増加するものとはなっていない．この意味により，ここで提案するものを「非単調連続DP」と呼ぶこととする．

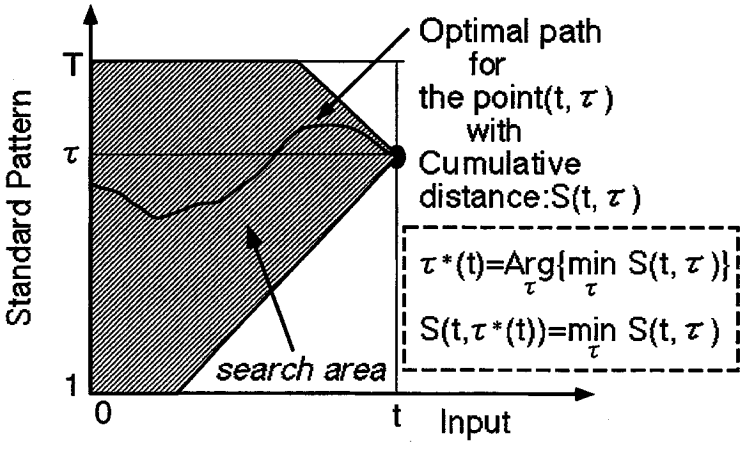


図 4.2 非単調連続DPにおけるパスの探索範囲

また，式(4.7)の$d(k,p(k))$に対する重み係数を$w(k)$とすると，重み係数$w(k)$の和は，

$$\sum_{k=0}^{t} w(k) = (1-\alpha)^t + \sum_{k=1}^{t} \alpha(1-\alpha)^{t-k} = 1 \tag{4.8}$$

となり，いかなるtにおいても重み係数$w(k)$の和が1に正規化された累積距離が得られることが分かる．実際，式(4.5)の重み係数の和が$\alpha+(1-\alpha)=1$となり，各フレーム毎に正規化されている．(このことは，正規化係数αが時間的に変化しても成り立つ.) これにより，累積距離の集合$\{S(t,\tau)|1 \leq \tau \leq T\}$内での比較が可能となる．

ここで，標準パターンがL個存在するとし，各パターンの累積距離を$S_\ell(t,\tau)(1 \leq \ell \leq L)$，しきい値を$h_\ell$，標準パターンのフレーム数を$T_\ell$とする．非単調連続DPの出力は，マッチングした標準パターンのカテゴリー番号$\ell^*(t)$とその標準パターン内でマッチングしたフレーム番号$\tau^*(t)$であり，

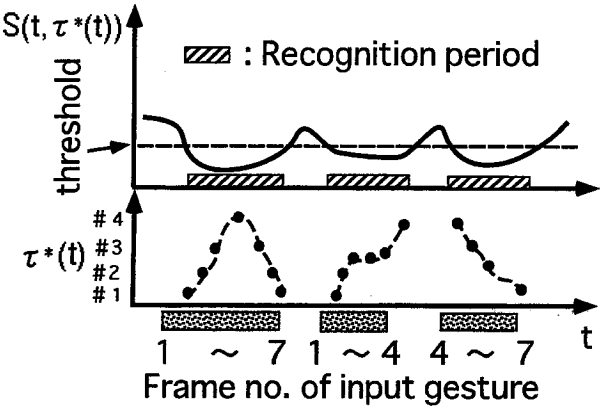
$$\{\ell^*(t),\tau^*(t)\} = \begin{cases} \mathrm{Arg}[\min_\ell\{\min_{1\leq\tau\leq T_\ell}(S_\ell(t,\tau)-h_\ell)\}] \\ \qquad \text{if } \exists\ell,\exists\tau \ \text{ so that } \ S_\ell(t,\tau) \leq h_\ell \\ \mathrm{null} \qquad \text{otherwise} \end{cases} \tag{4.9}$$

と表せる．ここで，Argは引数$\{\ell(t),\tau(t)\}$を返す関数，nullは空のカテゴリーを表す．また，このときの累積距離は，$S_{l^*(t)}(t,\tau^*(t))$と表せる．

ここで，本手法の特長を説明するため，1つ標準パターンを持ったジェスチャの認識システムを考える．この標準パターンとしては図4.3のような7枚のフレームからなるジェスチャ“手を挙げる”とする．

図 4.3 ジェスチャ“手を挙げる”のスナップショット

このとき，ジェスチャ“手を挙げる”のフレーム＃1～＃4を標準パターンとし，フレーム＃1～＃7，＃1～＃4，＃4～＃7の連続したジェスチャを入力する．すると，入力系列と標準パターンとの累積距離$S(t,\tau^*(t))$は，図4.4のように変化する．非単調連続DPでは，累積距離$S(t,\tau^*(t))$があるしきい値以下になった場合に，この標準パターンであると認識される．この図4.4では，標準パターンにジェスチャ“手を挙げる”のフレーム＃1～＃4しか含まれていないのにフレーム＃5～＃7も認識されている．これは，図4.3の＃1と＃7，＃2と＃6，＃3と＃5のフレーム同士が似ているためである．このように，非単調連続DPでは，図4.4のようにフレーム＃1～＃7のジェスチャを入力した場合だけでなく，戸惑っている動作（フレーム＃1～＃4，＃4～＃7）も1つの標準パターンで認識可能である．



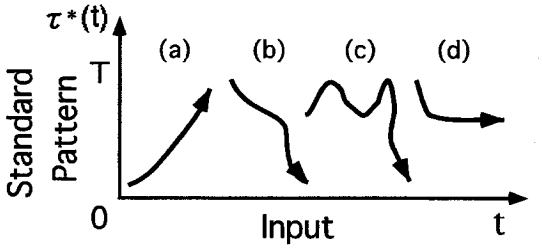
標準パターンは図4.3のフレーム# 1 ～# 4

図 4.4 非単調連続DPによる“手を挙げる”の部分変形動作3種についての認識

実は，このとき 非単調連続DPでは，ジェスチャ“手を挙げる”を認識しているのでは無く，“片手を上下に動かす”動作を認識していると言える．他にも，例として“両手を上下する”動作，“片手を左右に振る”動作などを認識するための標準パターンが考えられる．

複数の標準パターンがある場合，$S(t,\tau^*(t))$からしきい値を引いた値が最小となる標準パターン$l^*(t)$を認識結果とする．(ただし，この最小値が正の場合には認識結果は出力されない) これは，ジェスチャの大分類を行ったものと言える．

さらに，図4.4の下方のようにマッチングしたフレーム番号$\tau^*(t)$も得られ，この変化からジェスチャの細分類が可能となる．図4.5には，ある動作の部分的に変形した動作として，標準パターンに対して(a)順方向,(b)逆方向,(c)戸惑い,(d)静止した動作の4種類についての認識の様子を示した．これらの動作は，標準パターン中のどの部分区間で行われても認識できる．このような認識を従来の連続DPで行おうとすると，多くの標準パターンを用意する必要が生じ非効率的である．従って，このような認識が目的の場合には，非単調連続DPが有用である．



(a)順方向動作,(b)逆方向動作,(c)戸惑い動作,(d)静止した動作

図 4.5 非単調連続DPによる　ある動作の部分的に変形した動作4種の認識

また，本手法では戸惑い動作だけでなく，“少し大きい”,“非常に大きい”などの連続的な程度を示すジェスチャの認識も同じ枠組みで認識可能である．例えば，両手を広げると“大きい”,

表 4.1 半値フレーム数$k_{1/2}$と 正規化係数α

$k_{1/2}$	0	1	2	3	5
α	1	0.5	0.29	0.21	0.13
$k_{1/2}$	10	20	50	100	∞
α	0.07	0.03	0.01	0.007	0

狭めると“小さい”を示すものとする．このとき，標準パターンを両手を狭めた状態から広げるまでのフレーム区間とすればよい．これにより，大小の程度を連続的に認識することができる．このように，本手法は戸惑い動作だけでなく，程度を示す動作をも同じ枠組みで認識できるという特徴がある．

4.2.2 正規化係数αの時間可変化

$1 \leq k, 0 < \alpha \leq 1$のとき，重み係数$w(k)$は式(4.8)のように現時点$t$から過去になるに従って指数関数的に減少する．この重み係数の大きさが半減するまでのフレーム数として半値フレーム数$k_{1/2}$を

$$\frac{w(t-k_{1/2})}{w(t)} = (1-\alpha)^{k_{1/2}} = 0.5 \tag{4.10}$$

と定義する．このとき，

$$\alpha = 1 - (0.5)^{k_{1/2}^{-1}} \tag{4.11}$$

と，半値フレーム数$k_{1/2}$からαを決定できる．表4.1に半値フレーム数$k_{1/2}$とαとの例を記した．非単調連続DPでは，この半値フレーム数が大きいほど過去の履歴を用いた認識を行うことになる．また，半値フレーム数$k_{1/2}=0$，つまり$\alpha=1$のときは，現在のフレームのみを用いた1フレームマッチングとなる．

正規化係数αを時間的に一定とする場合，半値フレーム数が一定となり過去の一定時間分の履歴を用いた認識を行うことになる．しかし，特徴ベクトル系列が特徴空間内で描く軌跡上で考えると，一定距離の過去の情報を用いた認識とならない．ここで，特徴空間内にて特徴ベクトル系列の軌跡を過去にさかのぼったとき，重み係数$w(k)$が半減するまでの距離を「半値距離」と呼ぶこととする．この「半値距離」が，特徴ベクトルの時間変化量に比例する問題点について図4.6を用いて説明する．

いま，3つのジェスチャG1,G2,G3が，図4.6のように2次元の特徴ベクトル系列で記述されるものとする．入力ジェスチャは，ジェスチャG1のA点からB点，さらにC点へ移動するものを考える．もし，入力ジェスチャが特徴空間内でほぼ一定速度で移動したなら「半値距離」も一定となり，B点通過時にも過去の履歴を用いるためジェスチャG2,G3と混同することは無

いだろう．しかし，A点からB点に移動して静止するジェスチャの場合，静止時に「半値距離」が減少し，最終的には1フレーム間のマッチングと同様になる．つまり，特徴空間内での過去の履歴情報を失うことになる．従って，B点での静止動作時にジェスチャG2,G3と混同する可能性が生じる．ここで，半値フレーム数を大きくすれば「半値距離」も大きくなり，より長い時間において静止動作を識別できるであろう．しかし，半値フレーム数が大きいほど認識の遅れが大きくなり，通常の動作時に問題となる．

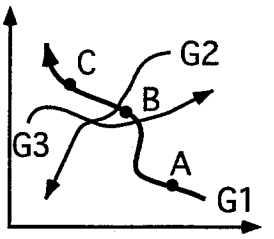


図 4.6　3つのジェスチャ(G1,G2,G3)の特徴特徴空間（2次元）上の軌跡

そこで，この問題の解決法の1つとして，半値フレーム数を特徴ベクトルの時間変化量に反比例させることで,「半値距離」を極力一定にすることが考えられる．ここで，特徴ベクトルの時間変化量$U(t)$を入力される特徴ベクトルu_tを用いて

$$U(t) = \sum_{k=1}^{N} \{u_t(k) - u_{t-1}(k)\}^2 \tag{4.12}$$

と表記し，半値フレーム数$k_{1/2}$を次のように時間的に変化させる．

$$k_{1/2}(t) = (\beta \cdot U(t))^{-1}. \tag{4.13}$$

βは特徴ベクトルの変化量$U(t)$を調整して適切な半値フレーム数$k_{1/2}(t)$を求めるための定数である．このとき式(4.11)より，$\alpha(t)$は，以下のように時間的に変化する．

$$\alpha(t) = 1 - (0.5)^{\beta \cdot U(t)} \tag{4.14}$$

さらに，この式を $\beta \cdot U(t) << 1$ として1次近似すれば，

$$\alpha(t) = \log 2 \cdot \beta \cdot U(t) = \beta' \cdot U(t) \tag{4.15}$$

と単純化できる．このように$\alpha(t)$を時間可変化することで「半値距離」を近似的に一定にできる．従って，静止動作を行ったときでも過去の履歴をもとに現在のジェスチャカテゴリーを認識することが可能となる．

4.3 戸惑い動作の認識

4.3.1 背景差分による低解像度特徴

非単調連続DPでは，先に述べたように静止したジェスチャも動的なものと同じ枠組みで扱うことができる．従って，従来用いていた時間差分[29]でなく，静止状態も検出可能な背景差分により特徴を抽出する．

初めに，サイズ$N_1 \times N_1$の入力画像$I(i,j,t)(0 \leq i,j < N_1, 0 \leq t)$と背景画像$I_{BG}(i,j)$から，次式により2値画像$I_b(i,j,t)$を求める．

$$I_b(i,j,t) = \begin{cases} 1 & \text{if } |I(i,j,t) - I_{BG}(i,j)| \geq h_c \\ 0 & \text{otherwise.} \end{cases} \tag{4.16}$$

ただし，h_cは画素値が変化したか決定するしきい値である．さらに，次式のように2値画像$I_b(i,j,t)$をサイズ$N_2 \times N_2$に縮小し，特徴ベクトル$f(k,l,t)(0 \leq k,l < N_2)$を求めた．

$$f(k,l,t) = \frac{1}{h^2} \sum_{0 \leq p,q < h} I_b(k \cdot h + p, l \cdot h + q, t). \tag{4.17}$$

ここで，pとqはともに整数，$h = N_1/N_2$である．この特徴ベクトルは，サイズ$N_2 \times N_2$に縮小した画像中の各領域内において，画素値が変化した割合，つまり，対象が移動したと推定される領域の割合である．

4.3.2 戸惑い動作の認識実験

実験装置として，SGI社のIndy(R4400 200MHz)と，付属のIndyComというカメラを用いた．実験は，オフィス内で椅子に座った1人の被験者に対して行った．カメラの視野は被験者のジェスチャが適切に入るように設定した．

CCDカメラの出力映像をAD変換して得られる画像は，サイズ160 × 120，1画素256階調のRGB画像であるが，認識には比較的輝度に強い影響を与えるグリーン成分のみを用いた．この画像を縮小しサイズ$N_1 = 64$として64 × 64の画像を特徴抽出部への入力とし，$N_2 = 3$として$3 \times 3 = 9$次元の特徴ベクトルを用いた．また，背景画像は被験者が両手を膝の上に置いているものとし，適宜更新した．

実験に用いたジェスチャは，(1)挙手（右手を挙げる），(2)バンザイ（両手を挙げる），(3)大小（両手にて，小さい$\longleftrightarrow$大きい），(4)高低（右手にて，低い$\longleftrightarrow$高い），(5)左右（右手にて，左へ$\longleftrightarrow$右へ）の5種類とした．括弧内はそれぞれの具体的な動作を示し，また，(3),(4),(5)は程度を示すジェスチャである．図4.7に各ジェスチャのスナップショットを示した．標準パターンは，図4.7下のように矢印の始点から終点までのジェスチャを撮影し特徴ベクトルを算出して作成した．例えば，ジェスチャ“挙手”の標準パターンは，図4.8(a)のように右手を下から上に挙げた部分のみを含んでいる．被験者は各動作を通常のスピードで行い，画像は15Hzでサ

ンプリングした．この実験で用いた標準パターンのフレーム長Tは10から27であった．また，式(4.16)のしきい値 h_cは人物の微小移動を考慮し30とした．

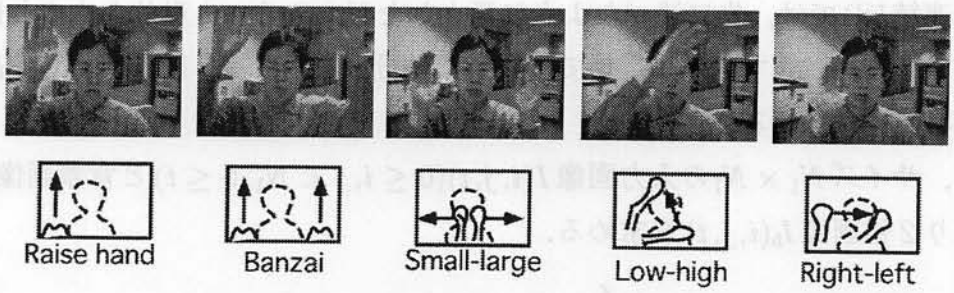


図 4.7 5種類のジェスチャのスナップショット

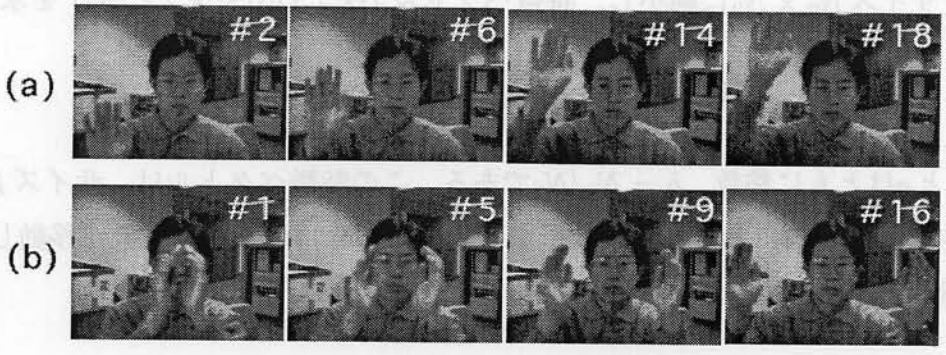
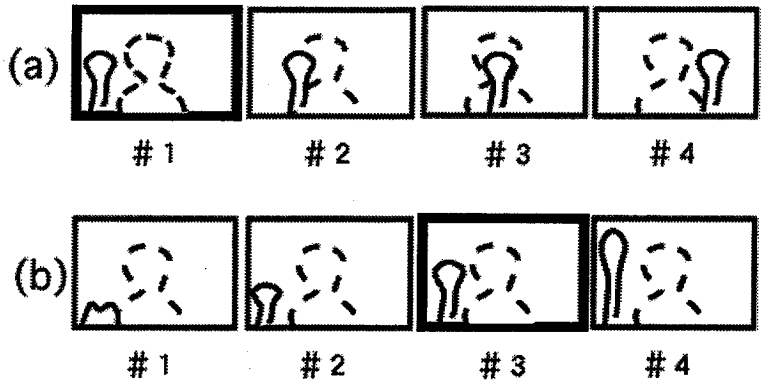


図 4.8 標準パターン (a)“挙手”,(b)“大小”

5種類のジェスチャには，2個の類似したフレームが存在する．すなわち，第1に“左右”の初期フレーム（図4.9(a)のフレーム#1）と“挙手”の中間フレーム（図4.9(b)のフレーム#3）とが類似しており，第2に“バンザイ”の中間フレームと“大小”の最終フレームとが類似している．

入力画像列は，5種類のジェスチャを順方向だけでなく逆方向，静止動作にて行い，特に類似フレームにて静止動作を行って撮影した．これは，ジェスチャ“挙手”と“バンザイ”では，戸惑った動作を想定し，“大小”,“高低”,“左右”の程度を示す動作では，連続的な程度の変化を認識させることを想定している．この入力画像列を人が認識した結果を図4.10(a)に示す．横軸は入力フレーム番号，縦軸は標準パターン中のフレーム番号$\tau^*(t)$である．入力画像列の全フレーム数は517であり，これを正規化係数$\alpha = \{1, 0.1, \text{可変}\}$の3通りに変化させて，以下で定義する認識率を求めた．

$$\text{認識率} = \frac{\ell^*(t)\text{と}\tau^*(t)\text{が正答したフレーム数}}{\text{入力したフレーム数}} \tag{4.18}$$



(a)“左右” のフレーム#1と (b)“挙手” のフレーム#3

図 4.9 標準パターン間の類似フレーム

表 4.2 認識結果

α	1	0.1	$\alpha(t)$
Recognition rate(%)	72	87	92

これは，各入力フレームにおいてマッチングした標準パターンの番号 $\ell^*(t)$ が入力したパターンの番号と一致し，かつ標準パターン内でマッチングしたフレーム番号$\tau^*(t)$ も正答する割合である．この実験では，フレーム番号$\tau^*(t)$ の許容誤差は± 2フレームとした．また，αが可変のときの式(4.15)のβ'の値は1とし，以下の式のようにαの最大値が1.0になるようにした．

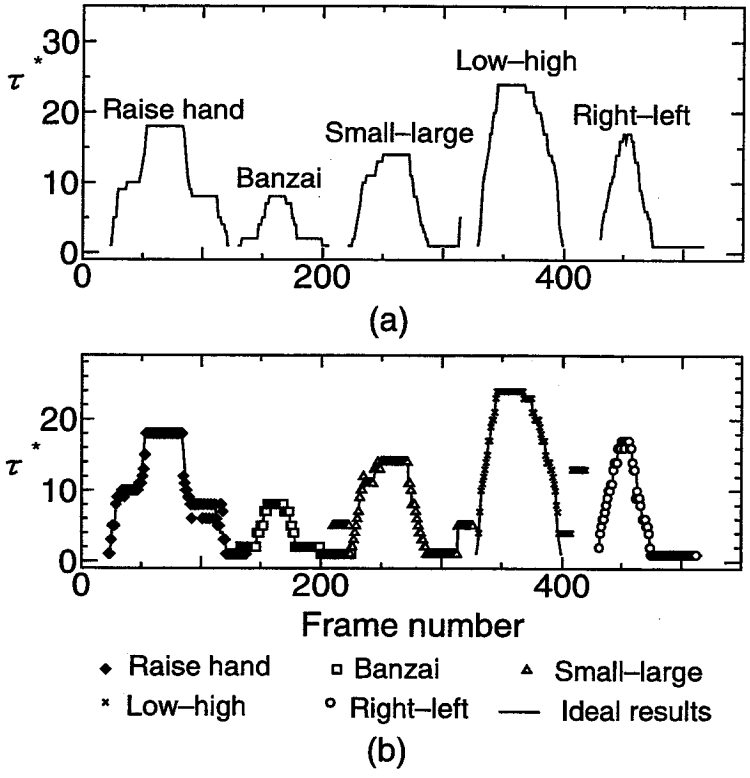
$$\alpha(t) = \min\{1.0, U(t)\}. \tag{4.19}$$

4.3.3 実験結果と考察

正規化係数αを変化させて認識実験を行った結果を表4.2に示す．α可変のときに約90%の認識率を示し，本手法の有効性が確認できた．α可変のときの誤認識の主な原因としては，ジェスチャ終了後もそのジェスチャであると誤認識したことが挙げられる．これは，標準パターンが，手が画面外へ移動する直前のフレームまで含んでいたために，手が画面外に出て特徴ベクトルが変化しなくなったときに直前の認識結果を保持してしまったためである．また，$\alpha = 1$のときは，認識が不安定で$\tau^*(t)$ に飛びが生じたことと静止動作時に他のジェスチャの類似フレームと誤検出したことが認識率低下の主な原因である．$\alpha = 0.1$のときには認識の安定性があったものの，ジェスチャ開始時の認識に遅れが生じていた．

図4.10(b) に，α可変のときのスポッティング認識結果を示す．図4.10(b) の菱形の各点は，“挙手” 動作であると認識された結果であり，横軸が入力フレーム番号，縦軸が標準パターン中で

マッチングしたフレーム番号$\tau^*(t)$とを示している．正方形，三角，×印，丸印のそれぞれも同様に“バンザイ”，“大小”，“高低”，“左右”と認識された結果を示している．また，実線は人が認識した理想的な結果である．この図4.10(b)から非単調連続DPによるスポッティング認識により，戸惑い動作だけでなく，程度を示すジェスチャをも連続的に認識できていることがわかる．



(a)視察による認識結果,(b)非単調連続DPによる認識結果

図 4.10 認識実験結果

図4.11に正規化係数αを変化させたときの認識の様子を示す．ここでの入力ジェスチャは，ジェスチャ“左右”において，右手を右から左に移動して，さらに右に移動して静止したものである．このとき，図4.9に説明したようにジェスチャ“挙手”と類似フレームとなる．図4.11の白丸は“左右”，菱形は“挙手”と認識された結果であり，実線が人による理想的な認識結果である．

図4.11(a)に示すように，過去の情報を用いない場合($\alpha = 1$)は，すぐに混乱が生じている．しかし，図4.11(b)に示すように過去の情報を多く用いること($\alpha = 0.1$)で，しばらくは前の軌跡情報を維持できた．ただし，過去の情報を多く用いることによってジェスチャ開始時の認識に時間遅れが生じている．

そこで，式(4.15)で示したようにαを時間可変とした結果，図4.11(c)に示すように認識の時間遅れが小さくなり，また，混乱も生じなくなった．

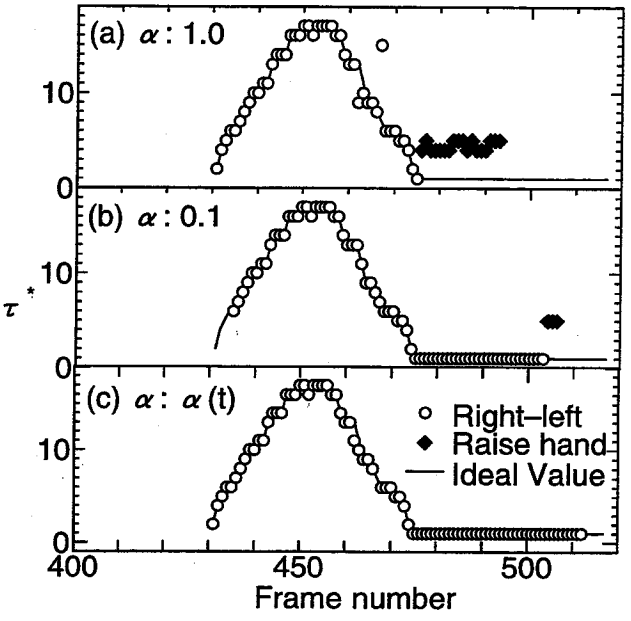


図 4.11 正規化係数αの影響

4.3.4 実時間ジェスチャ認識システム

Indy を 1 台用い図4.12のような実時間認識システムを作成した．標準パターンとしたジェスチャは，認識実験のものと同様である．認識結果はジェスチャ名に続いて stop,normal,reverse と表示し，それぞれ，静止時，順方向時，逆方向時とした．実時間での認識実験を行なった結果，背景画像の更新を適宜行えば，約8割の認識率が得らることが分かった．

4.4 移動ロボットの自己位置推定

4.4.1 非単調連続DPの必要性

伊藤らは，カメラの揺れに対処するために 1 枚の入力画像から抽出する情報量を必要最低限に押え，この特徴を複数画像分用いることで，逆にロバストな位置推定が可能であることを示した[75]．ここでは，ロボットが地図作成時と異なるスピードで走行してもマッチング可能とするために，入力パターンを時間的に伸縮させる非線形マッチングを行なっている．

複数画像の情報を用いて位置推定を行なう研究として，松本ら[69]は，画像間の距離があるしきい値以下の位置を候補とし，ロボットが走行するに従い，この候補位置を絞り込んでいく方法を用いている．また，Zheng[72]らは， DP を用いて大局的位置を推定した．この方法では，非線形マッチング可能であるが，始点と終点が既知である必要があるため，ロボットは地図と同じ経路を走行しなくてはならない．

一方，伊藤らはロボットが地図の任意の部分区間を走行しても位置推定ができるよう，連

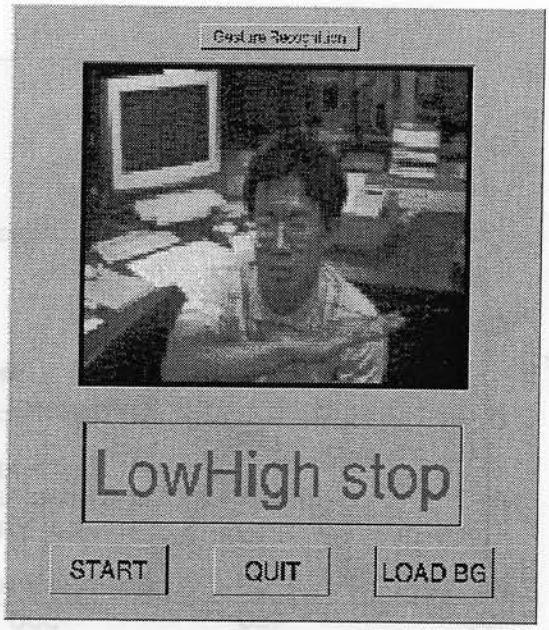


図 4.12 実時間ジェスチャ認識システム

続DP[15]を応用したReference Interval-Free 連続DP[60] を用いて位置を推定している[75]. Reference Interval-Free 連続DPは，音声認識で提案された手法であり，入力フレームと同期して地図中の任意の部分区間長のフレーム列をスポッティング的に検出することができるため，地図中の部分区間の走行時や走行速度の変化時でも位置推定可能である．実際，入力パターンが，時間軸方向に対して順方向(monotonic)に$\frac{1}{2}$倍以上2倍以下の伸縮があってもマッチング可能である．しかし，位置推定の場合，入力パターンの伸縮はロボットが逆方向に走行することもあるため，例えば–2〜2倍の伸縮を想定する必要がある．つまり，ロボットの走行方向に依存せずに位置推定を行なうためには，順方向と逆方向のマッチングを同時に行なうことが可能な手法が望まれる．4.2節で提案した'非単調連続DP' は，このような目的に最適である．そこで，非単調連続DPをロボットの位置推定用のマッチング手法として導入する．

4.4.2 従来の特徴量の問題点

環境撮影時の視野としては，通常のカメラ程度のもの[69],[75]と全方位を捉えたもの(panoramic view)とに大別できる．前者の通常の視野を用いて大局的な位置を推定する場合は，地図作成時と入力時のカメラ方向を一致させる必要があるため，カメラ方向の制御が必要である．これに対して，全方位を視野に持つ全方位視覚センサを用いれば，このような制御が不要となる．Zhengら[72]は，視野として進行方向と異なる方向の垂直スリットを用いた．各位置における情報は，このスリット内の画素値をRGB毎に積算するためカメラの垂直方向の揺れにロバストである．ただし，カメラの水平方向や回転方向の揺れの影響が懸念される．また，Facchinettiら[67]は天井の画像系列を用いた位置推定を行っているが，我々の想定している環境では，水

平方向の視野の方がより多くの情報を有している．Bangら[73]は全方位センサを用いてロボット誘導を行なっているが，過去の履歴が既知であることを仮定し，大局的位置の推定を行わずに直接局所位置を推定している．前田ら[68]は，画像サイズが，垂直方向64画素，水平方向（周方向）256画素の全方位画像を用いている．そして，この画像の各水平ラインをフーリエ変換し，この低周波数側の強度成分（64 × 32画素）を用いて位置推定している．

一方，我々が用いていた従来の特徴抽出法[75]では，通常のカメラを用いていたが，低解像度な特徴を複数画像分用いてマッチングすることでロバストな位置推定を実現している．そこで，本節では，全方位視野を たかだか5分割して特徴を求める方法を導入する．

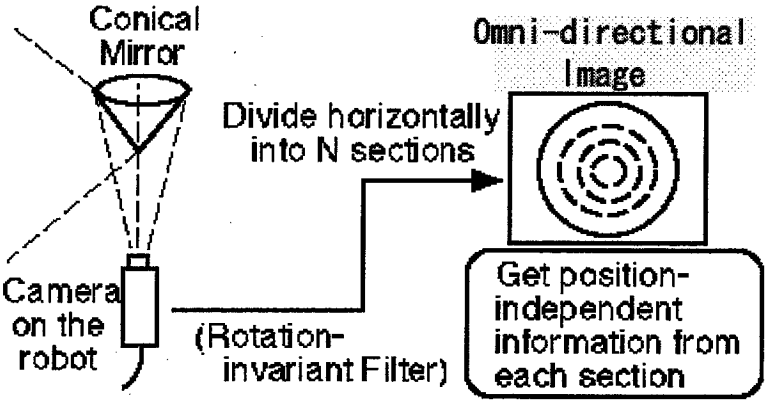


図 4.13 回転不変な特徴抽出

4.4.3 回転不変な特徴量の概要と定式化

全方位視野を水平にN分割すると，図4.13 に示すように，全方位画像上では同心円で区分されるN個のドーナツ状の領域に分けられる．次に全方位画像全体において局所的なK次元の特徴を抽出し，これを各ドーナツ状の領域内において平均化したものを特徴量とする．この特徴量は，$N \times K$次元の特徴ベクトルであり，位置が一定であればロボットがどの方向を向いていても一定である．これが，本節で導入する回転不変な特徴量である．局所的な特徴としては，画素値そのものやラプラシアンフィルタなどの近似的に等方性の成り立つものが考えられる．また，特徴ベクトルの次元数$N \times K$はたかだか5×12くらいを想定しており，ロボット走行時の揺れにもロバストとなるだけでなく，地図情報のメモリ量が低減できる．

前節で述べた回転不変特徴の一例として，本手法の評価実験にて用いた特徴抽出法を定式化する．

全方位画像(x, y)中のi番めのドーナツ状の領域を$R_i (i = 1, \cdots, N)$と記述し，以下の式で定義する．

$$\{(x, y) | r_{i-1}^2 \leq (x - x_c)^2 + (y - y_c)^2 < r_i^2\} \tag{4.20}$$

ここで，(x_c, y_c)は全方位画像の中心座標であり，この点での輝度値は全方位視野における垂直

表 4.3 特徴抽出のためのRGB空間の分割（明るさ）
j(j=1,⋯,16):特徴番号, V_{RGB}:明るさ, thr_{black}:黒のしきい値, thr_{mid}:中間の明るさのしきい値, thr_{white}:白のしきい値

j	intensity	condition
1	black	$V_{RGB} < \mathrm{thr}_{black}$
2,⋯,8	dark	$\mathrm{thr}_{black} \leq V_{RGB} < \mathrm{thr}_{mid}$
9,⋯,15	light	$\mathrm{thr}_{mid} \leq V_{RGB} < \mathrm{thr}_{white}$
16	white	$\mathrm{thr}_{white} \leq V_{RGB}$

下方向のものである．この点を中心にして半径r_{i-1}以上r_i未満のドーナツ状の領域がR_iである．この半径r_iは以下の式で求める．

$$r_i = \frac{i}{N}(r_{max} - r_{min}) + r_{min} \tag{4.21}$$

ここで，全方位画像の最小半径をr_{min}，最大半径をr_{max}とした．

次に，各分割領域R_i内から局所的な情報を有効に抽出する手法を示す．従来法[75]では，RGB値を独立に平均した3次元の特徴量を用いていた．しかし，視野の広い全方位画像でこの特徴量を求めても，位置による特徴量の変化が小さくなる．そこで，より詳細な情報を得るため，以下で述べる16色の色の出現頻度を用いる．領域R_i内の各画素値のRGB値をV_R, V_G, V_B，これらの平均（明るさ）を$V_{RGB} = (V_R + V_G + V_B)/3$とする．このとき，表4.3,4.4のようにRGB空間を16個の領域$C_j(j = 1, \cdots, 16$，以下jを特徴番号と呼ぶ)に分割した．表4.3では，3個のしきい値(thr_{black},thr_{mid},thr_{white})を用いて明るさを4分（黒，暗い，明るい，白）していることを示している．このうち，“暗い”と“明るい”については更に表4.4の様にしきい値thr_dを用いて7分割し，色情報を取り出している．

ここで，入力された全方位画像中のi番めの領域R_i内のj番めの色領域C_jに含まれるピクセル数を$K_{ij}(i = 1, \cdots, N, j = 1, \cdots, 16)$，$i$番めの領域$R_i$内のピクセル総数を$A_i$とし，回転不変な特徴ベクトル$f_{ij}$を以下の式で求める．

$$f_{ij} = \frac{K_{ij}}{A_i \cdot H_j} \tag{4.22}$$

ただし，地図に含まれるすべての画像について色領域C_jの出現比率H_jを求め，このH_jで正規化した．

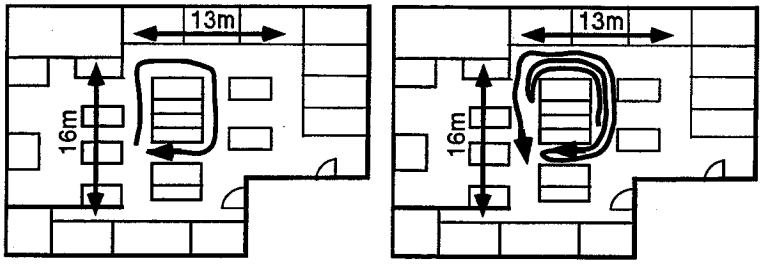
4.4.4 移動ロボットを用いた実験

今回提案した位置推定手法の評価を行うため，山澤ら[70]が提案した全方位視覚センサHyperOmni Visionを移動ロボットNomadに設置した．実験データは，フロアがカーペット敷

表 4.4 特徴抽出のためのRGB空間の分割（色）
thr_d:差があると判断するしきい値

j	color	condition
2 or 9	R	$V_R - V_G > \mathrm{thr}_d$ and $V_R - V_B > \mathrm{thr}_d$
3 or 10	G	$V_G - V_R > \mathrm{thr}_d$ and $V_G - V_B > \mathrm{thr}_d$
4 or 11	B	$V_B - V_R > \mathrm{thr}_d$ and $V_B - V_G > \mathrm{thr}_d$
5 or 12	RG	$V_R - V_B > \mathrm{thr}_d$ and $V_G - V_B > \mathrm{thr}_d$
6 or 13	GB	$V_G - V_R > \mathrm{thr}_d$ and $V_B - V_R > \mathrm{thr}_d$
7 or 14	BR	$V_B - V_G > \mathrm{thr}_d$ and $V_R - V_G > \mathrm{thr}_d$
8 or 15	grey	$V_X - V_Y \leq \mathrm{thr}_d$ for all $X, Y \in \{R, G, B\}$

きのオフィス内の通路をリモートコントロールでロボットを走行させて採集した．このとき，人はロボットから約2m離れていた．



(a) 地図作成時およびS1,S2,S3の場合の経路，(b) 速度変化時（S4,S5）の経路．

図 4.14 実験で用いた画像系列の取得経路

地図作成時は，図4.14(a)に示すような経路の内側寄りを約20cm／秒の速度で走行し，2 frame/s のフレームレートで160 × 120pixelの大きさの地図画像系列を作成した．本実験を行った環境の様子を，図4.15に示す．観葉植物や本棚，会議机などがあり比較的変化に富んだ環境といえる．入力としては，次の五つの場合を設定し時系列画像を作成した．

[S1] 地図作成時と同一経路をほぼ同一の速度で走行した場合

[S2] 地図作成時の約25cm外側の経路を同一速度で走行した場合

[S3] 図作成時の約50cm外側の経路を同一速度で走行した場合

[S4] 図作成時と同一経路内を約-1.8～1.8倍の速度で走行した場合

[S5] 図作成時の0～50cm外側の経路内を約-1.8～1.8倍の速度で走行した場合

ここで，S4,S5の経路は図4.14(b)に示すようになっており，地図作成時に走行した経路を順

方向，逆方向に走行し随時停止している．

更に，地図作成時の経路と異なった経路を約-1.8～1.8倍の速度で走行した場合(地図外)の時系列画像を作成した．各画像系列のフレーム数は，地図が304フレーム，S1～S5の場合の入力がそれぞれ 319, 320, 356, 879, 986フレーム，地図外の場合が276フレームとなった．

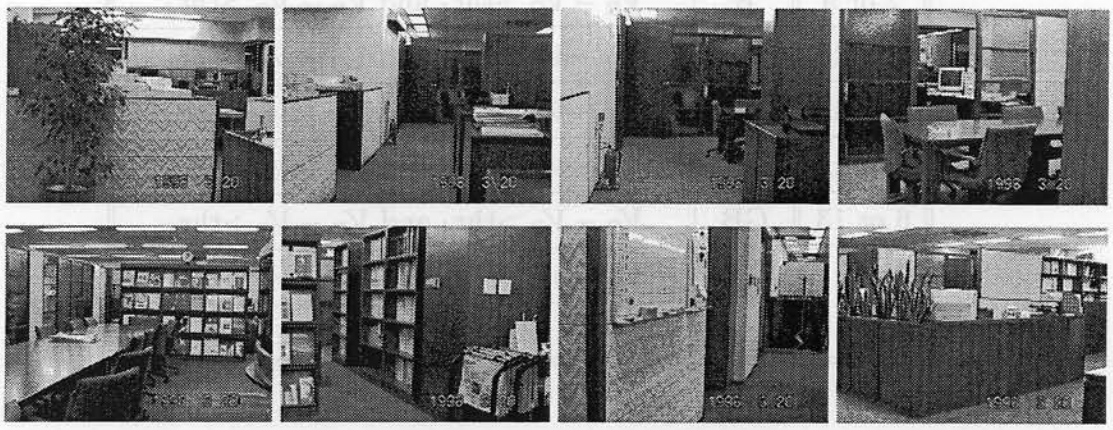


図 4.15 実験環境の様子

4.4.3節で述べた回転不変特徴は，画像中心が(78, 58)，半径が15～55pixel の領域から抽出した．また，色抽出のためのしきい値は，経験的に$\mathrm{thr}_{black} = 35$,$\mathrm{thr}_{mid} = 100$, $\mathrm{thr}_{white} = 230$,$\mathrm{thr}_{d} = 15$とした．さらに，事前に地図画像系列中の色の出現頻度を求めたところ，色特徴番号6,13,7,14は出現頻度が非常に小さかった．このため，位置推定時にはこの色を除いた12色の特徴を用いた．非単調連続DPの局所距離$d(t, \tau)$は，地図画像系列および入力画像から求めた特徴ベクトルの対応する要素の差の2乗和で求めた．また，傾斜パターンとしては，図4.1(b)のように−2～2倍の伸縮を許すものとした．

評価実験では，S1～S5のそれぞれの場合について地図外の画像系列を付加した入力画像系列を作成し，正答したフレーム数を入力フレーム総数で割って位置推定率を求めた．正答かどうかの判断は，地図内走行時は，推定結果が正しいフレーム番号の±10フレーム以内である場合を正答とし，地図外では，地図外と推定した場合を正答とした．この正しいフレーム番号は，入力画像に対応する地図画像のフレーム番号を人間が目視で対応付けして求めた．また，式(4.9)のしきい値hは，大きすぎると地図外走行時に地図内にいると誤推定することが多くなり，小さすぎると地図内走行時に地図外にいると誤ることが多くなる．従って，本実験では，しきい値hを変化させ，最も高い位置推定率を採用した．

さらに，S1～S5のそれぞれの場合について，分割数を$N = 1, 2, \cdots, 8$，正規化係数を$\alpha = 1.0, 0.3, 0.1, 0.03, 0.01$と変化させて位置推定率を求め，それぞれのパラメータの影響を調べた．

4.4.5 実験結果と考察

図4.16～4.20に，それぞれS1～S5の場合について分割数Nと正規化係数αを変化させたときの位置推定率を示した．初めに，分割数Nについて検討する．S1～S5のすべての場合において，分割数Nが小さい（$N=1,2$）と推定率が低く，$N=3,4,5$で最も高くなる傾向がある．これは，Nが小さいと一枚の画像から得られる情報量が少ないためと考えられるが，複数フレームを用いる（$\alpha<1.0$の場合）ことで推定率が大きく向上している．このことは，$N=3,4,5$の場合よりも$N=1,2$の場合の方が顕著である．Nが6以上において推定率が低下する原因としては，視野を細分化し過ぎたためにロボット走行時のカメラの揺れや位置ずれによって各領域の視野が変化したためと考えられる．

正規化係数αに関しては，$\alpha>0.1$（表1より半値フレーム数が7フレーム未満）の場合に推定率が低下している．これは，マッチングに影響するフレーム数が少なく，局所的に類似した場所の判別が困難となるためである．$\alpha=0.1,0.03$において推定率が最大となっているが，$\alpha=0.01$では幾分低下している．これは，αが小さすぎると時間分解能が減少して地図内から地図外への移動時などに過去の影響を受けて誤推定するためである．

横にずれた経路を走行した場合(S2,S3)の位置推定率は，地図と同一経路で同一速度の場合(S1)に比べ，1～5%程度低下している．しかし，50cm程度の横ズレに対しても最高97%の推定率を示しており，本手法のロバスト性が示せた．さらに，速度を変化させた場合(S4,S5)でも，約95%の推定率を示しており，本手法が，逆方向走行時や静止時に対応できることを示せた．

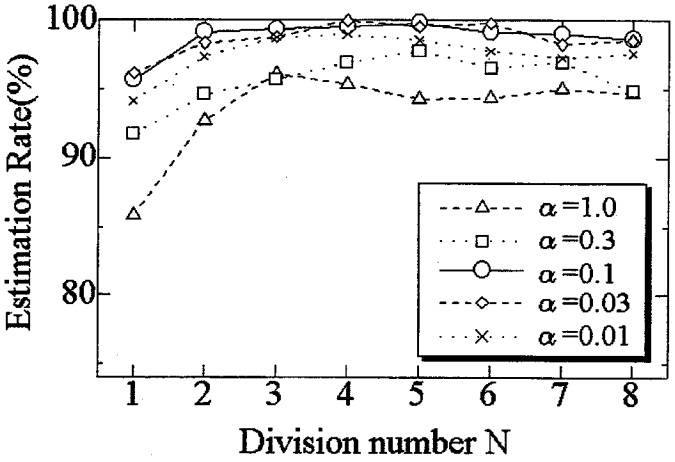


図 4.16 位置推定率（[S1]同一経路，同一速度）

図4.21には，位置ずれ，速度変化がともにある場合(S5)の$N=4,\alpha=0.1$における位置推定結果を示した．図4.21(a)は人手で対応づけした正しいフレーム番号，図4.21(b)は位置推定結果（地図画像のフレーム番号）である．この図から逆方向走行時や停止時でも位置推定可能で

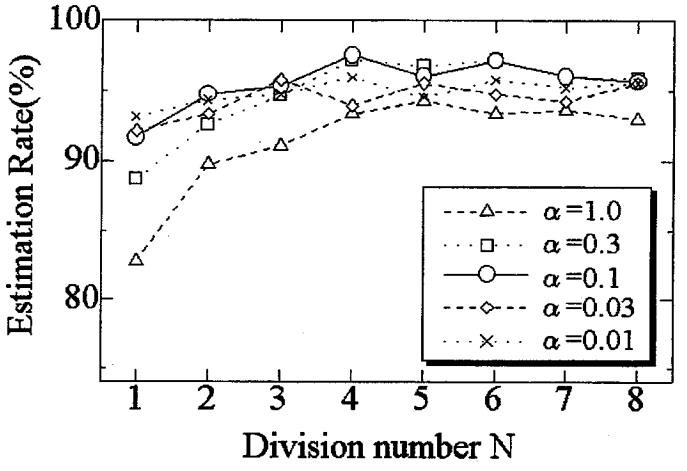


図 4.17 位置推定率（[S2]横ズレ25cm，同一速度）

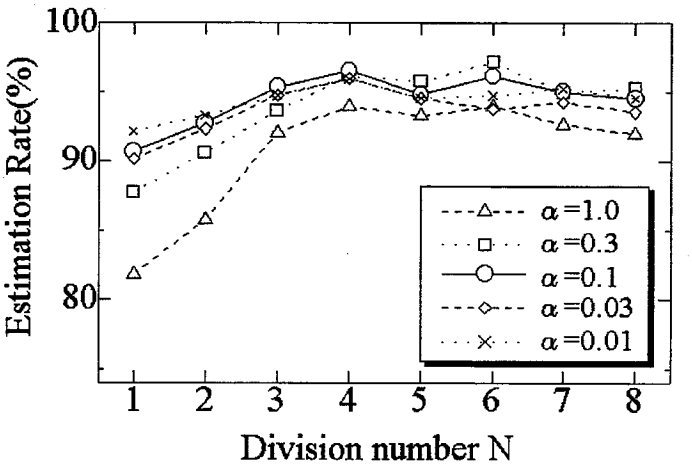


図 4.18 位置推定率（[S3]横ズレ50cm，同一速度）

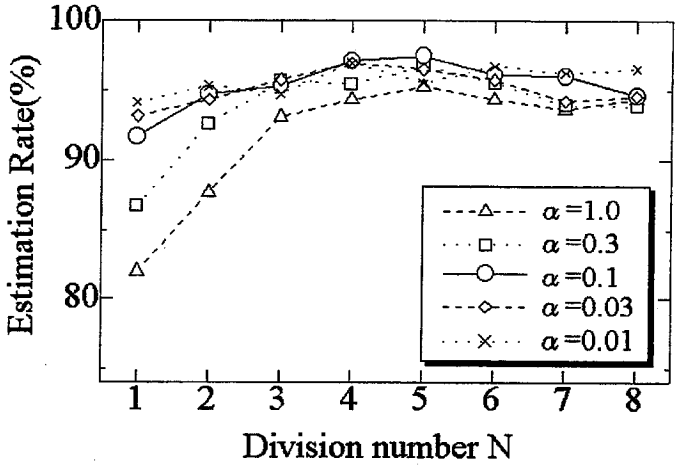


図 4.19 位置推定率（[S4] 同一経路，速度変化）

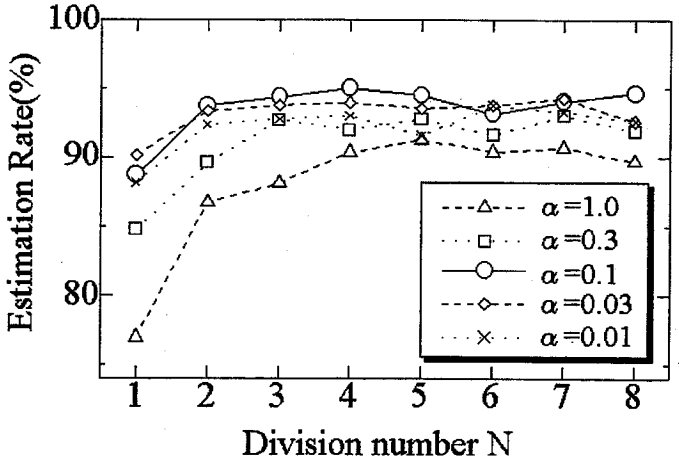
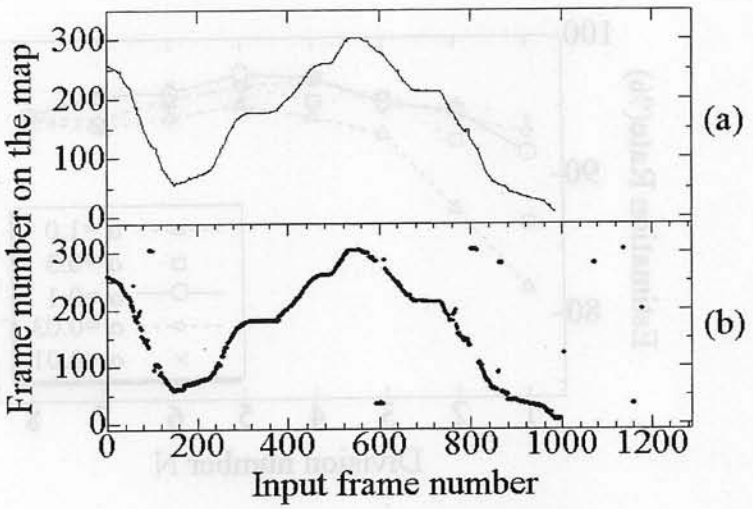
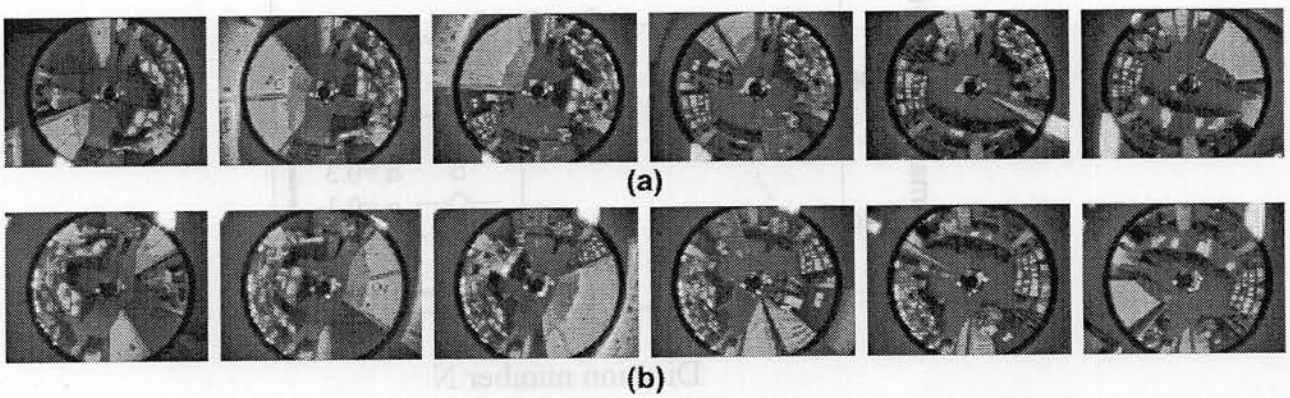


図 4.20 位置推定率（[S5] 横ズレ 0～50cm，速度変化）



(a) 人手による正しい対応（フレーム番号），(b) 位置推定結果（分割数 $N = 4$, 正規化係数 $\alpha = 0.1$）.

図 4.21 S5 の場合の位置推定結果



(a) 入力画像系列（610〜660 フレーム，10 フレーム毎）(b) 推定結果（対応する地図画像）.

図 4.22 位置推定の様子

あることが分かる．さらに，図4.22には，位置推定の様子をHyperOmni Visionの画像で表示した．図4.22(a)が入力画像，図4.22(b)が推定された地図画像系列中の画像である．この図から，カメラ方向に依存せず位置を推定できていることが分かる．

4.5 まとめ

本章では，連続DPに非単調性を導入した非単調連続DPを提案した．また，この非単調連続DPにより，戸惑っている動作や程度を示すジェスチャのスポッティング認識が可能であることを実験にて示した．さらに，正規化係数αを時間的に変化させて「半値距離」を一定にする手法の有効性を示した．

問題点としては，特徴抽出法において背景差分を行なっているため，人物の位置が変化すると認識率が悪化することである．従って，今後の課題としては，距離画像や色情報を用いるなどして，よりロバストな特徴抽出法を開発することである．

また，地図中の複数の画像特徴系列とのマッチング法として，ロボットが地図作成時と逆方向に走行したときや停止時でもスポッティング的に位置推定を可能とするために，非単調連続DPを導入した．さらに，ロボットの大局的な位置推定をロバストに行わせるために有効な，水平に分割した全方位視野の各領域内の局所的な特徴の平均値を用いた．この特徴量はロボットの回転に不変なため，地図作成時と異なる方向であっても位置推定が可能である．移動ロボットを用いた実験によって，カメラ方向が異なりかつ逆方向走行時や停止時の場合でも位置推定が可能であることを示した．

今後の課題としては，（1）本手法を用いた実時間誘導システムの実現，（2）非単調連続DPを逐次行ないながら地図画像のトポロジーを自動的に構築する手法の提案，などが挙げられる．この（2）では，随時，局所パスに経路の接続情報をもたせることでトポロジー地図を構築したいと考えている．

第 5 章

重み減衰型RIFCDP

5.1 はじめに

近年，膨大なマルチメディアデータの検索や要約，モーダル変換，構造解析を実現できる時系列パターンの検索技術が，重要な研究課題となっている．このような状況において，2つの時系列データ間について，任意の長さを持ち，かつ互いに類似した区間（類似区間）対を検出することを可能とする，Reference Interval-free 連続DP(RIFCDP) [60] が提案されている．

このRIFCDPは，標準パターンを入力パターンと同一にして類似区間を検出するIncremental RIFCDP(IRIFCDP)[61]に拡張され，実時間の音声要約や話題境界検出が実現されている．更に，4.4.1節で述べたように移動ロボットの位置推定にも適用されている．

しかし，従来のRIFCDPでは標準パターン中の各フレームにおいて，検出したい類似区間長分の累積距離および入力時刻を保持・算出しているため，メモリ量と計算量が非常に大きくなる．例えば，のべ1日分の画像データベースからRIFCDPを用いて検索するためには5Gバイトものメモリ量が必要となり実用上困難な事態を生じている．

そこで，本章では，局所距離にかかる重みを，標準パターンにおいて過去に溯るに従って指数関数的に減少させて累積距離を計算する考えを導入することにより，計算量とメモリ量の軽減を可能とし，かつほぼ類似の機能をもつ重み減衰型 RIFCDP を提案する．この手法は，類似区間長分すべての累積距離および入力時刻を保持・算出するRIFCDPに対し，わずか過去2フレームの累積距離から漸化的に現時点の累積距離を算出する．本手法を用いると，のべ1日分の画像データから検索する場合でも約100Mバイトと，従来のRIFCDPの約50分の1のメモリ量でシステムが実現できる．また，計算量についても従来の約1/16となる．

本章の構成は，5.2節で重み減衰型 RIFCDP を提案し，その概念や特徴を詳細する．また，5.3節にて本手法で用いている重み減衰型ウインドウについて説明する．さらに，5.4節にて，これまでRIFCDPが適用されていなかったジェスチャ動画像 [29] を用いて本手法の有用性を示す．これにより，今まで音声の分野で主に用いられていた検索や要約，話題境界検出技術を，人物の行動理解，手話認識さらにはテレビ映像などのマルチメディアを扱う分野へ適用できることを示す．

5.2 重み減衰型 RIFCDP

5.2.1 RIFCDP の問題点

従来の RIFCDP は，各時刻において（1）連続 DP[15] の計算，（2）累積距離履歴などのコピー及び更新，（3）整合度計算，（4）類似区間検出を行なっている．これを，図 5.1 を用いて順に説明する．図 5.1 の縦軸は T フレームの標準パターン，横軸が入力であり，現時刻を t，標準パターン中の注目フレームを τ とする．また，RIFCDP では，検出したい類似区間の最小フレーム数 N_{min} と最大フレーム数 N_{max} を，時系列データ中の類似区間長の分布から予め決定しておく．

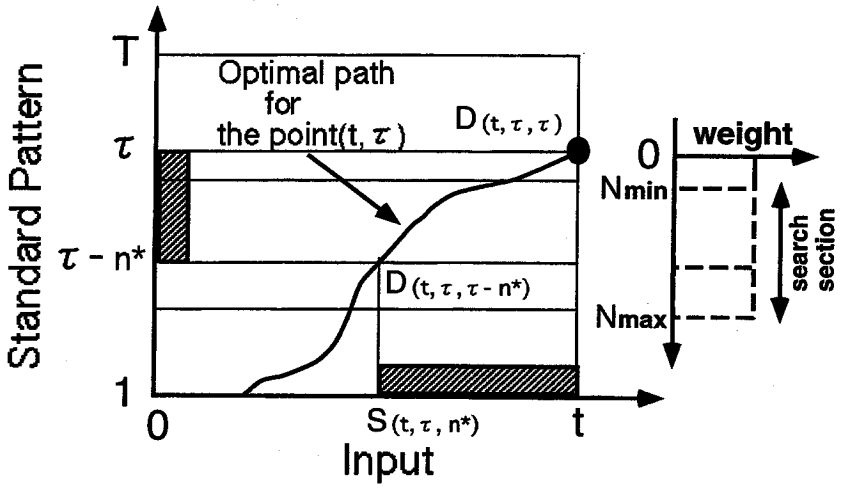


図 5.1 従来の RIFCDP

（1）連続 DP の計算

連続 DP により点 (t,τ) を終点とする最適パスを求める．ただし，これは標準パターン $1\sim\tau$ と入力との最適パスである．

（2）累積距離履歴などのコピー及び更新

点 (t,τ) を終点とする最適パス上の区間 $\tau - N_{max}\sim\tau$ において累積距離 $D(t,\tau,\tau-n)(0 \leq n < N_{max})$ および入力時刻 $S(t,\tau,n)$ を保持し，それぞれのコピーおよび更新を行なう．ここで，n は標準パターンの軸上で点 (t,τ) から遡ったフレーム数である．

（3）整合度計算

点 (t,τ) を終点とする最適な類似区間を求めるために以下の処理を行なう．まず，最適パス上の $\tau - N_{max}\sim\tau - N_{min}$ において，類似区間が $(\tau-n,\tau):(S(t,\tau,n),t)$ と仮定した場合の整合度 $A(t,\tau,n)$ を次のように求める．

$$A(t,\tau,n) = \frac{D(t,\tau,\tau) - D(t,\tau,\tau-n)}{n} \qquad (N_{min}-1 \leq n < N_{max}) \tag{5.1}$$

ここでは，最適パス上の点 (t,τ) までの累積距離から点 $(S(t,\tau,n),\tau-n)$ までの累積距離を引き，これを n で正規化して区間 $(\tau-n,\tau):(S(t,\tau,n),t)$ の整合度としている．次に，最適な類

似区間の始点を決定するために，以下の式でn^*を求める．

$$n^* = \mathrm{Arg} \min_{N_{min}-1 \leq n < N_{max}} A(t, \tau, n). \tag{5.2}$$

ここで，点(t, τ)を終点とする最適な類似区間は$(\tau - n^*, \tau) : (S(t, \tau, n^*), t)$，そのときの整合度は$A(t, \tau, n^*)$と求まる．

（4）類似区間検出

$\tau = N_{min}, \cdots, T$について，整合度$A(t, \tau, n^*)$がしきい値$D_{thr}$以下の場合に既に得られた類似区間と処理（3）で得られた区間とをマージする．

以上がRIFCDPで行なわれる処理である．ここで，点(t, τ)を終点とするパス上の局所距離$d(t, \tau - n)$に対する重みを$w(n)$，この重みの系列を“ウインドウ”と呼ぶこととする．処理（2）と処理（3）では，図5.1の右方に示すように，均一な重みのウインドウの長さをN_{min}からN_{max}まで変化させて最適な類似区間を算出していることになる．つまり，$N_{max} - N_{min}$個のウインドウを用いて点(t, τ)を終点とする最適な類似区間を求めていることになる．ここでは次の2つの問題がある．

[問題1] 計算量とメモリ量が大きい．

[問題2] 最適性原理を満たしていない．

[問題2] は，処理（1）の「連続DPの計算」の部分で述べたように，点(t, τ)を終点とする最適パスが，標準パターン1〜τと入力との最適パスであって，標準パターン$\tau - n^*$〜τとのそれではないために生じた問題である．しかし，この [問題2] は5.4節の評価実験で示すように，大きな問題とはなっていない．[問題1] の原因は，処理（2）で最適パス上の区間$\tau - N_{max}$〜τのすべてにおいて累積距離などの情報を保持・更新していることと，処理（3）で区間$\tau - N_{max}$〜$\tau - N_{min}$において式（5.1）を計算しているためである．

表5.1，表5.2の上段にRIFCDPの計算量とメモリ量（バイト）の概算を示した．ここで，Nは時系列データの次元数，1変数4バイト，C^+, C^cをそれぞれ加減乗除，比較やコピーの計算量とした．また，$C^+ \doteqdot C^c, N >> 1, T >>$ [類似区間数] と仮定して概算値を示した．更に，局所距離はユークリッド距離の自乗とした．例えば，のべ1日分の画像データ（$T = 2.5 \times 10^6$フレーム）から抽出された時系列特徴データから，最大フレーム数$N_{max} = 90$フレーム（3秒）で類似区間検出する場合を考える．時系列データの次元数を$N = 10$として表5.2から算出すると，$(40 + 24 \times 90) \times 2.5 \times 10^6 \doteqdot 5$Gバイトと莫大なメモリ量が必要となる．また，計算量については5.2.2節にて重み減衰型RIFCDPと比較して検討する．

5.2.2 重み減衰型RIFCDPの概念

重み減衰型RIFCDPでは，各時刻において局所距離にかかる重みを，過去に溯るに従って指数関数的に減少させて累積距離を計算する連続DPによる整合度計算（5.2.1節の（1）の部分），類似区間検出（5.2.1節の（4）の部分）の2つの処理を行なうだけでよい．つまり，従

表 5.1 RIFCDP と重み減衰型 RIFCDP の計算量

	(1)CDP	(2)copy	(3)matching	(4)merge
RIFCDP	$3NTC^{+}$	$\underline{2N_{max}TC^{c}}$	$\underline{3N_{max}TC^{+}}$	0
wd-RIFCDP	$3NTC^{+}$	0	0	0

表 5.2 RIFCDP と重み減衰型 RIFCDP のメモリ量（バイト）

	(1)CDP	(2)copy	(3)matching	(4)merge
RIFCDP	4NT	$\underline{24N_{max}T}$	0	0
wd-RIFCDP	4NT	0	0	0

来の RIFCDP（5.2.1節）で行なっていた（2）と（3）の処理を無くし，図5.2の右方に示すように指数関数的に重みを減少させた1つの近似ウインドウ（重み減衰型ウインドウ）を局所距離にかけて整合度を算出している．従来の RIFCDP での（2）と（3）の処理では，均一な重みのウインドウの長さを変化させているため (図5.1の右方を参照)，最適な類似区間を検出できる反面，計算量やメモリ量が増大していた．

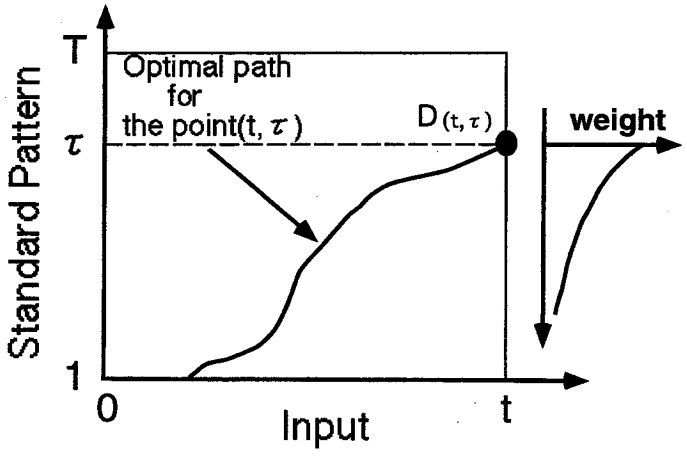


図 5.2 重み減衰型 RIFCDP

しかし，重み減衰型 RIFCDP では，1つのウインドウしか用いないため類似区間の情報が得られず，累積距離 $D(t,\tau)$ のみが得られ，これがそのまま整合度になっている．そこで，類似区間検出処理では，累積距離 $D(t,\tau)$ において最小フレーム数以上の長さの右肩上がりの谷線を類似区間として検出する．図5.3の上方に $D(t,\tau)$ を画像として表示したときの，実際の類似区間 (t,τ^{*}) と $D(t,\tau)$ があるしきい値以下になる領域を表示した．図5.3の下方には実際の類似

区間(t,τ^*)上における累積距離$D(t,\tau^*)$の変化とそのときの類似区間検出結果の概念図を示した．累積距離$D(t,\tau^*)$は，実際の類似区間が始まると減少し始め，区間終了直後に上昇し始める．本手法の共通区間検出法では，この累積距離$D(t,\tau^*)$があるしきい値以下の場合に検出するため図5.3の下方の斜線部のように実際の類似区間と比べて若干始点・終点が遅れて検出される．

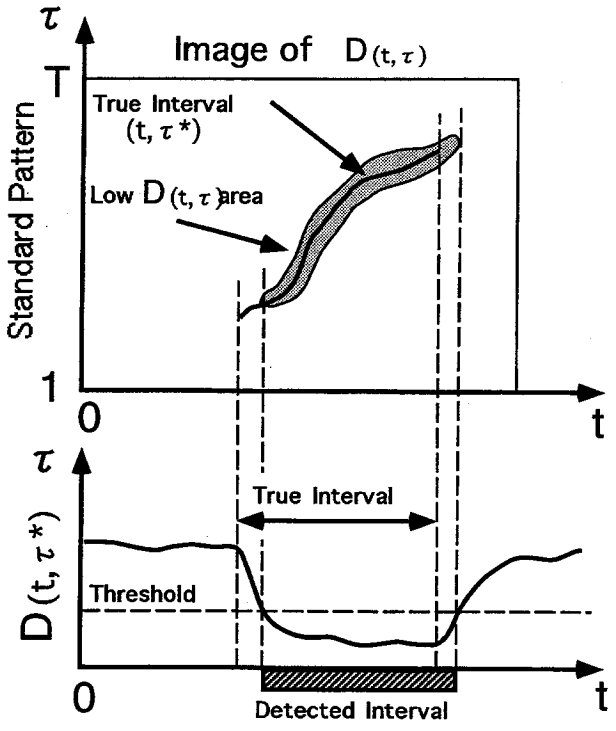


図 5.3 類似区間の検出

ここで，従来のRIFCDPと比較したときの重み減衰型RIFCDPの利点と欠点をまとめる．

[利点] 計算量とメモリ量が少ない．

[欠点1] 類似区間が若干遅れて検出される．

[欠点2] 近似ウインドウによって検出率が低下する．

また，[欠点1]については従来のRIFCDPと比べて2～3フレーム程度の遅れとなった．更に，5.3節にて[欠点2]に関する理論的な考察を示した．

表5.1，表5.2の下段に重み減衰型RIFCDPの計算量とメモリ量を概算した．5.2.1節の例で取り上げた，のべ1日分の画像データに提案手法を適用する場合，メモリ量を概算すると$40\times 2.5\times 10^6 = 100$Mバイトと従来のRIFCDPの約50分の1となる．また，計算量については，$C^+ \doteqdot C^c$と仮定すると重み減衰型RIFCDPでは従来のRIFCDPの約1/16になる．

5.2.3 定式化

一つの標準パターン Zは，標準動作をとらえたTフレームの動画像から得られる特徴ベクトル z_τの系列

$$Z = \{z_\tau | 1 \leq \tau \leq T\} \tag{5.3}$$

で表す．ここで，特徴ベクトル z_τはその次元数を Nとすると

$$z_\tau = (z_\tau(1), z_\tau(2), \cdots, z_\tau(N)) \tag{5.4}$$

である．入力画像からも同様な特徴ベクトル系列 $u_t(0 \leq t < \infty)$が連続的に得られる．このとき，u_tと z_τとの局所距離を $d(t,\tau)$と表記する．ここで，入力，標準パターンの時間軸をそれぞれt,τと区別する．

更に，点(t,τ)を終点とした標準パターンと入力系列との累積距離を $D(t,\tau)$で表す．重み減衰型RIFDPでは$D(t,\tau)$を以下のような漸化式で更新する．

初期条件 $(t=0)$：

$$D(0,\tau) = d(0,\tau). \quad (1 \leq \tau \leq T) \tag{5.5}$$

初期条件 $(t=1)$：

$$D(1,1) = d(1,1) \tag{5.6}$$

$$D(1,2) = \alpha \cdot d(1,2) + (1-\alpha) \cdot D(0,1) \tag{5.7}$$

$$D(1,\tau) = \min \begin{cases} \alpha \cdot d(1,\tau) + (1-a) \cdot D(0,\tau-1) \\ a \cdot d(1,\tau) + b \cdot d(1,\tau-1) + (1-a-b) \cdot D(0,\tau-2). \end{cases} \quad (3 \leq \tau \leq T) \tag{5.8}$$

漸化式 $(2 \leq t)$:

$$D(t,1) = d(t,1) \tag{5.9}$$

$$D(t,2) = \min \begin{cases} \frac{\alpha}{2} d(t,2) + \frac{\alpha}{2} d(t-1,2) + (1-\alpha) \cdot D(t-2,1) \\ \alpha \cdot d(t,2) + (1-\alpha) \cdot D(t-1,1) \end{cases} \tag{5.10}$$

$$D(t,\tau) = \min \begin{cases} \frac{\alpha}{2} d(t,\tau) + \frac{\alpha}{2} d(t-1,\tau) + (1-\alpha) \cdot D(t-2,\tau-1) \\ \alpha \cdot d(t,\tau) + (1-\alpha) \cdot D(t-1,\tau-1) \\ a \cdot d(1,\tau) + b \cdot d(t,\tau-1) + (1-a-b) \cdot D(t-1,\tau-2). \end{cases} \quad (3 \leq \tau \leq T) \tag{5.11}$$

ここで，$\alpha(0<\alpha\leq 1)$ は正規化係数，

$$a:b:(1-a-b)=1:(1-\alpha):(1-\alpha)^2 \tag{5.12}$$

である．ここでは，図5.4のような傾斜パターンを採用している．

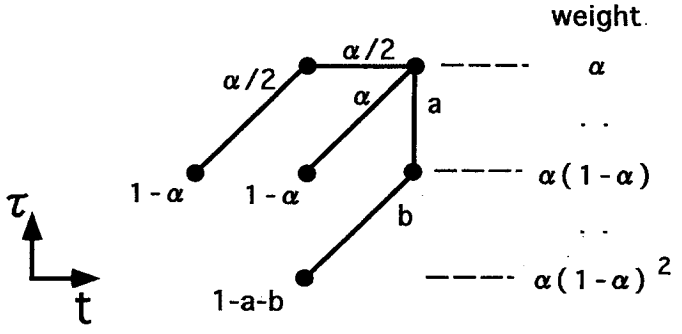


図 5.4 重み減衰型RIFCDPの傾斜パターン

この重み減衰型RIFCDPでの $d(t,\tau-n)(0\leq n<\tau)$ に対する重み係数を $w(n)$ とすると，

$$w(n)=\begin{cases}\alpha(1-\alpha)^n & (0\leq n<\tau-1)\\ (1-\alpha)^n & (n=\tau-1)\end{cases} \tag{5.13}$$

となる．この重みは，過去に溯るに従って指数関数的に減少しており，これを重み減衰型ウインドウと呼ぶ．また，重み係数 $w(n)(0\leq n<\tau)$ の和は常時1と正規化されている．実際，漸化式(5.10)(5.11)の重み係数の総和が $\frac{\alpha}{2}+\frac{\alpha}{2}+(1-\alpha)=1, \alpha+(1-\alpha)=1$ または $a+b+(1-a-b)=1$ となり，各フレーム毎に正規化されている．つまり，1つの重み減衰型ウインドウによって，各入力時刻 t において累積距離の集合 $\{D(t,\tau)|1\leq\tau\leq T\}$ 内での比較が可能となる．

5.2.4 類似区間検出法

5.2.2節で述べたように重み減衰型RIFCDPにおける類似区間の検出法では，累積距離 $D(t,\tau)$ において最小フレーム数以上の長さの右肩上がりの谷線を類似区間として検出するればよい．このとき，入力フレームに対してフレームワイズ，つまり逐次的に検出できれば，Incremental RIFCDP(IRIFCDP)[61]に拡張しやすい．このIRIFCDPはRIFCDPにおいて入力パターンを標準パターンに逐次コピーしながら類似区間を検出する手法であり，実時間の音声要約や話題境界検出が実現されている．従って，本節では，これらのことを踏まえた類似区間検出法の1例を紹介する．

入力時刻 $t-1$ までに類似区間とみなした J 個の区間を $(\tau_s(j),\tau_e(j)):(t_s(j),t_e(j))(1\leq j\leq J)$ とし，$t-1$ において接続された標準パターン軸上の区間を $(\tau_s(j)^{t-1},\tau_e(j)^{t-1})$ とする．このとき，時刻 t において以下の（1）～（5）の処理を実行する．

（1）$\tau = 1, \cdots, T$について，累積距離$D(t,\tau)$があるしきい値D_{thr}以下の区間(τ_s^t, τ_e^t)のすべてにおいて（2）～（5）を実行．

（2）$j = 1, \cdots, J$について，（3）を実行．

（3）区間(τ_s^t, τ_e^t)が，区間$(\tau_s(j)^{t-1}, \tau_e(j)^{t-1})$に対して傾きが1/2～2で接続しているとき，

$$\begin{aligned} \tau_e(j) = \tau_e^t, \quad t_e(j) = t \\ \tau_s(j)^t = \tau_s^t, \quad \tau_e(j)^t = \tau_e^t \end{aligned} \tag{5.14}$$

として，区間(τ_s^t, τ_e^t)をj番めの類似区間に追加する．

（4）$j = 1, \cdots, J$のいずれも（3）を満たさなければ，Jを1増加させ，以下の式で区間(τ_s^t, τ_e^t)を新しい類似区間として追加する．

$$\begin{aligned} \tau_s(j) = \tau_s^t, \quad \tau_e(j) = \tau_e^t, t_s(j) = t_e(j) = t \\ \tau_s(j)^t = \tau_s^t, \qquad\qquad \tau_e(j)^t = \tau_e^t \end{aligned} \tag{5.15}$$

（5）$j = 1, \cdots, J$について，以下の2つの条件を満たすj番目の類似区間を，最小フレーム数N_{min}より短く，これ以上成長することが無い区間として削除する．

$$\begin{aligned} \tau_e(j) - \tau_s(j) < N_{min} \\ t_e(j) < t - 2. \end{aligned} \tag{5.16}$$

5.3 重み減衰型ウインドウの特性

この節では，提案手法で用いた重み減衰型ウインドウの特性を，従来のRIFCDPで用いている重みが一定なウインドウと比較して説明する．

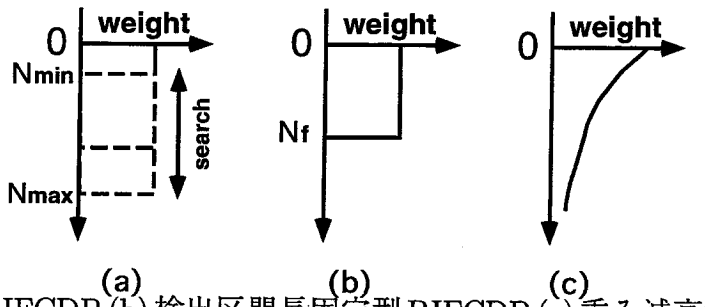
ある時系列データの中に区間長xの類似区間が存在する場合，区間内の各局所距離に対する重みが一定である理想ウインドウ$w_x^d(k)$が検出に最適であると仮定する．

$$w_x^d(k) = \begin{cases} \frac{1}{x} & (1 \le k \le x) \\ 0 & (x < k). \end{cases} \tag{5.17}$$

ここで，添字dは理想的(desirable)であることを示し，kは式（5.1）で用いた標準パターン軸の$n+1$に対応する変数である．また，重み$w_x^d(k)(1 \le k)$の総和は1に正規化されている．

一方，従来のRIFCDPでは，図5.5(a)に示すように長さLの類似区間を検出対象とする一定重みウインドウ$w_L^c(k)$（添字のcは一定(constant)を意味する）

$$w_L^c(k) = \begin{cases} \frac{1}{L} & (1 \le k \le L) \\ 0 & (L < k) \end{cases} \tag{5.18}$$



(a)従来のRIFCDP,(b)検出区間長固定型RIFCDP,(c)重み減衰型RIFCDP

図 5.5 各種ウインドウ

を用い，更にLをN_{min}からN_{max}まで変化させて整合度を算出している．このため，$N_{min} \leq x \leq N_{max}$であれば理想ウインドウ$w_x^d(k)$と完全に一致する一定重みウインドウ$w_L^c(k)(L=x)$が存在し，類似区間を最適に検出できる．これが，従来のRIFCDPの特色であるが，先に述べたように計算量とメモリ量が増大する．

そこで，この問題を解決するために，若干の検出力低下を覚悟の上で図5.5(b)のように1つの一定重みウインドウ$w_{N_f}^c(k)(N_{min} < N_f < N_{max})$のみを用いることが考えられる（検出区間長固定型RIFCDP）．この場合，検出区間長N_fは，想定したすべての類似区間の検出率を高くするように設定すれば良いが，計算量・メモリ量ともに累積距離履歴などのコピー及び更新（表5.1，表5.2の(2)copy参照）の負荷が大きいまま残る．

これに対し，今回提案した重み減衰型RIFCDPでは，図5.5(c)のように1つの重み減衰型ウインドウを用いることで，従来の連続DPとほぼ同じ計算量とメモリ量で整合度を求めることができる．この重み減衰型ウインドウ$w_\alpha^w(k)$は

$$w_\alpha^w(k) = \alpha(1-\alpha)^{k-1} \qquad (1 \leq k) \tag{5.19}$$

と表記できる．ここで，添字wは重み減衰型(weight decreasing)であることを，添字αは正規化係数を示す．しかし，この重み減衰型ウインドウを用いることによって類似区間検出率がある程度低下するはずである．

ここで，検出対象の類似区間に最適な理想ウインドウが$w_x^d(k)$の場合に，あるウインドウ$w_q^p(k)$（$(p,q) = \{(c,L),(w,\alpha)\}$）を用いて類似区間を検出したときの検出率を類似区間検出率$J_q^p(x)$とし，以下の式で定義する．

$$J_q^p(x) = \sum_{k=1}^{\infty} \min(w_x^d(k), w_q^p(k)). \tag{5.20}$$

これは，実際に用いたウインドウと区間長xの理想ウインドウとの一致度を求めたものであり，$0 \leq J_q^p(x) \leq 1$である．このとき，(p,q)は，重み一定型ウインドウの場合では(c,L)，重み減衰型ウインドウの場合では(w,α)とする．式(5.17),(5.18),(5.19)より重み一定型ウインドウの類似区間検出率$J_L^c(x)$は，

$$J_L^c(x) = \begin{cases} \frac{x}{L} & (1 \le x \le L) \\ \frac{L}{x} & (L < x) \end{cases} \tag{5.21}$$

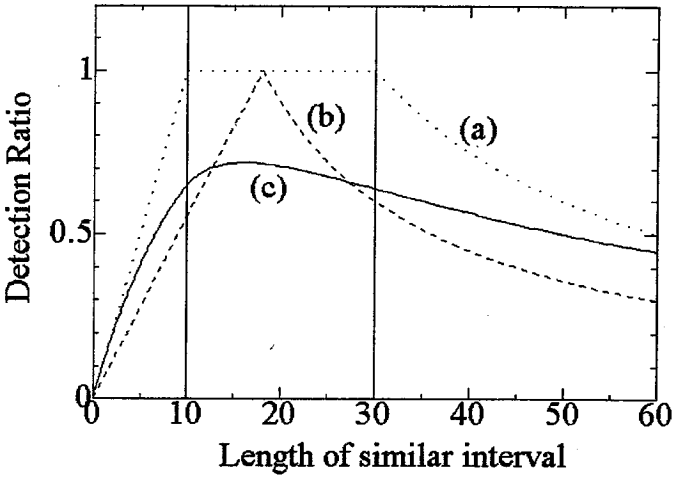
となり，重み減衰型の類似区間検出率 $J_\alpha^w(x)$ は，

$$J_\alpha^w(x) = \begin{cases} 1-(1-\alpha)^x & (1 \le x \le \frac{1}{\alpha}) \\ \frac{\log(\alpha x)}{x\log(1-\alpha)} - (1-\alpha)^x + \frac{1}{\alpha x} & (\frac{1}{\alpha} < x) \end{cases} \tag{5.22}$$

と求まる．また，従来のRIFCDPにおける類似区間検出率 $J^R(x)$ は式(5.21)より

$$J^R(x) = \max_{N_{min} \le L \le N_{max}} J_L^c(x) = \begin{cases} \frac{x}{N_{min}} & (1 \le x < N_{min}) \\ 1 & (N_{min} \le x \le N_{max}) \\ \frac{N_{max}}{x} & (N_{max} < x) \end{cases} \tag{5.23}$$

と求まる．



(a) 従来のRIFCDP,(b) 検出区間長固定型RIFCDP,(c) 重み減衰型RIFCDP

図 5.6 類似区間検出率

検出したい類似区間の区間長が10～30フレームである場合に $N_{min} = 10, N_{max} = 30, N_f = 18, \alpha = 0.1$ と設定し，従来のRIFCDP，検出区間固定型RIFCDP，重み減衰型RIFCDPの類似区間検出率 $J^R(x), J_{N_f}^c(x), J_\alpha^w(x)$ を図5.6に示した．横軸が検出したい類似区間の区間長 x である．図5.6(a)のRIFCDPでは，N_{min} から N_{max} までの類似区間検出率が1であるのに対し，図5.6(b)の検出区間長固定型RIFCDPでは，$x = N_f$ をピークとして大きく低下する．また，図5.6(c)の重み減衰型RIFCDPでは，N_{min} から N_{max} まで類似区間検出率が0.7程度と低いものの，N_{min}～N_{max} にてほぼ一定であるという特徴がある．

ここで，3種のウインドウの特徴を表5.3にまとめた．類似区間検出率は N_{min} から N_{max} までの平均値を示した．また，計算量およびメモリ量は，表5.1および表5.2を参照して概算した．

表 5.3　3種のウインドウの特徴
D.R.:平均した類似区間検出率, M.A.:メモリ量, C.B.:計算量

	D.R.	M.A.	C.B.
RIFCDP	1.0	$24N_{max}T$	$5N_{max}TC^{+}$
Fixed RIFCDP	0.8	$24N_fT$	$2N_{max}TC^{+}$
wd-RIFCDP	0.7	$4NT$	$3NTC^{+}$

通常，$N_{max} >> N, N_{max} \doteqdot N_f$であることを考えると，重み減衰型RIFCDPは，検出率の低下をほどほどに押えながら計算量およびメモリ量を大きく低下させていると言える．また，5.4節の評価実験で示すように，5.2.4節で新たに提案した類似区間検出法により，この検出率の低下もほとんど無くすことができる．

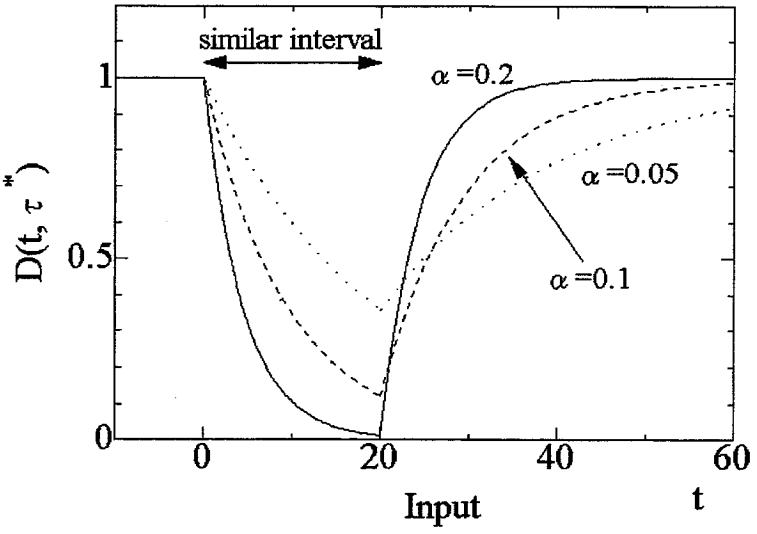


図 5.7　類似区間検出シミュレーション

次に，重み減衰型RIFCDPにおける累積距離変化を図5.7に示す．この図は，理想的な類似区間(ここでは区間長20フレーム)が入力の0〜20フレームにおいて存在した時の累積距離変化をシミュレーションしたものであり，図5.3の下方の図に対応する．この図5.7から，$\alpha > 0.1$であれば類似区間が始まると同時に急速に累積距離が減少し始め，類似区間の終了とともに急速に累積距離が上昇し始めることが分かる．従って，5.2.4節で示した類似区間検出法にてしきい値を最適に決定すれば検出遅れを小さくできると予想される．

ここで，検出したい類似区間長が得られたときに，正規化係数α をいかに決定すれば良いかが問題となる．そこで，重み減衰型RIFCDPの正規化係数αを変化させた場合の類似区間検出

率$J_\alpha^w(x)$を図5.8に示した．この図でも横軸が検出したい類似区間の区間長xである．この図5.8から，検出したい類似区間長の分布が分かれば類似区間検出率を最大にするように正規化係数αを決定できる．

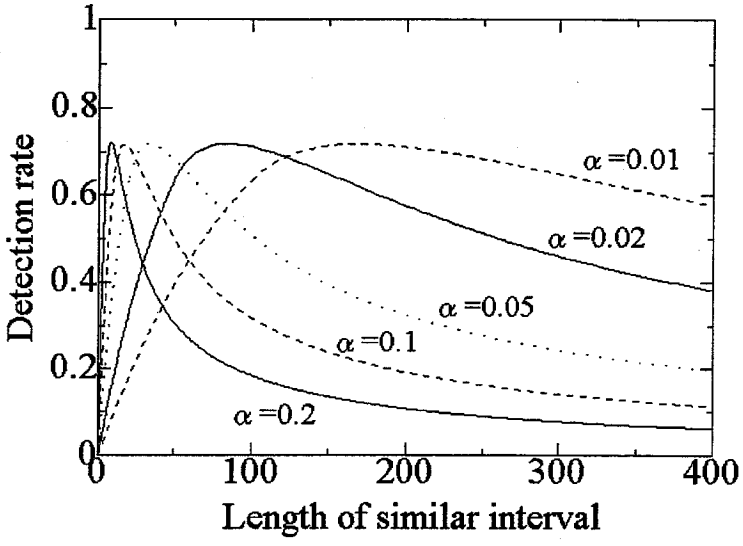


図 5.8 重み減衰型RIFCDPの類似区間検出率

5.4 ジェスチャ動画像の検索

5.4.1 実験方法

本手法の評価を行なうため，ジェスチャ動画像データを採集した．実験装置として，SGI社のIndy(R4400 200MHz)と，付属のIndyComというカメラを用いた．実験は，オフィス内で椅子に座った1人の被験者に対して行った．1枚の画像中の人物領域を3×3に分割し，各領域中の変化領域の割合を求め，2.2.2節で示した9次元の低解像度特徴を抽出した．また，入力系列u_tと標準パターンz_τとの局所距離$d(t,\tau)$は以下の式で求めた．

$$d(t,\tau)=1-(\sum_{k=1}^{N}u_t(k)\cdot z_\tau(k))^2\{(\sum_{k=1}^{N}u_t(k)^2)(\sum_{k=1}^{N}z_\tau(k)^2)+N\sigma^4\}^{-1}. \tag{5.24}$$

ここで，σは，信号パワーが小さい場合に局所距離を増大し，マッチング対象外とみなすための定数である．このσを適切に決定すれば，ジェスチャ動画像では静止時を，音声では無音区間を検出対象外にできる．今回の実験では，$\sigma=0.002$とした．

実験に用いたジェスチャは，(1)ばんざい（両手），(2)こちらへ（右手），(3)バイバイ（右手），(4)いいえ（右手），(5)まる（両手），(6)左へ（左手），(7)手をたたく（両手），(8)右へ（右手）の8種類である．図5.9に各ジェスチャのスナップショットを示す．被験者は各動作を通常のスピードで行い，画像は15Hzでサンプリングした．

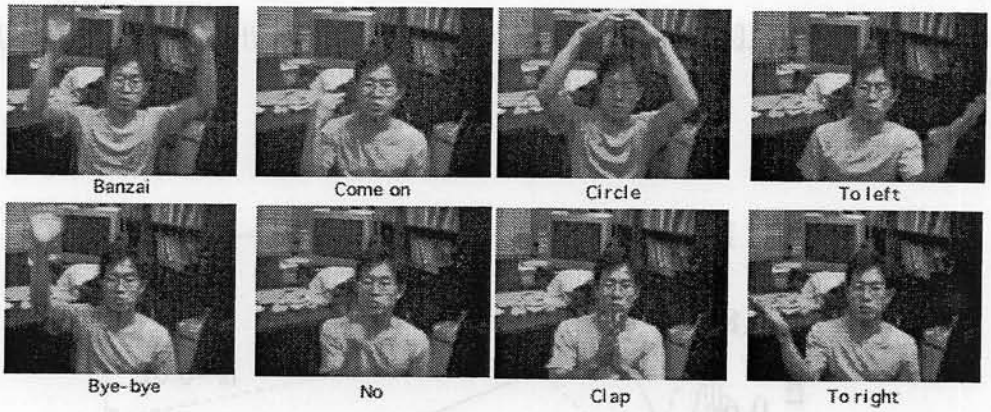


図 5.9 8種類のジェスチャのスナップショット

実験では，標準パターンは(1)～(8)のジェスチャを順に行なったものを用い，入力系列は，逆の順序で(8)～(1)のジェスチャを行なったものを用いた．入力系列としては以下の2種を用意した．

[S1] 標準パターンと衣服および背景が同一なもの

[S2] 標準パターンと衣服および背景が異なるもの

これらのデータ中の各ジェスチャのフレーム数は 17～29だったため，類似区間の最小フレーム数を $N_{min} = 10$，最大フレーム数を $N_{max} = 30$ と設定した．実験結果を示す検出率は以下の式で定義し，

$$\text{検出率} = \frac{\text{類似区間} \cap \text{検出区間}}{\text{類似区間} \cup \text{検出区間}} \tag{5.25}$$

RIFCDP，重み減衰型RIFCDPのそれぞれについて，この検出率を最大にするよう整合度および累積距離のしきい値 D_{thr} を変化させた．また，重み減衰型RIFCDPの正規化係数αは，N_{min}, N_{max} の値を考慮し検出率が最大となるよう図5.8を参照して$\alpha = 0.1$と決定した．

5.4.2 実験結果

表5.4に実験結果を示す．この結果から重み減衰型RIFCDPは，従来のRIFCDPと同等の検出率を達成できていることがわかる．また，衣服・背景が異なる場合［S2］で検出率が50%程度に落ちているが，これは音声検索に適用した先の報告[60]における実験結果と検出率の定義を同一にして比較したところ同程度であった．つまり，衣服・背景が同じ場合［S1］は，音声の分野であり得ないほど理想的な状態となっていたと考えられる．

図5.10には，衣服・背景が同じ場合［S1］における類似区間と2手法の検出結果を示した．縦軸が標準パターン，横軸が入力である．ジェスチャ(2)こちらへ（右手）と(8)右へ（右手）が紛らわしく，2手法とも誤検出している．検出遅れは，RIFCDPにて1～3フレーム，重み減衰型RIFCDPでは，2～6フレームであった．

図5.11には，衣服・背景が同じ場合［S1］における局所距離および整合度を画像（縦軸：標

表 5.4 実験結果

D.R.:検出率，［S1］:標準パターンと入力で衣服・背景は同じ，［S2］:標準パターンと入力で衣服・背景は異なる

	D.R. [S1]	D.R. [S2]
RIFCDP	74.4%	42.5%
wd-RIFCDP	69.4%	50.2%

準パターン，横軸：入力）として表示した．いずれの値も 0 ～ 1 に正規化されているため，0 は黒く 1 を白く表示した．図5.11(a) は局所距離画像であるが，ノイズが多くこの情報からは類似区間をほとんど見分けることができない．ここでも，ジェスチャ(2) こちらへ（右手）と (8) 右へ（右手）が似通っているため (2) と (8) の交差領域が黒く（局所距離が近く）なっている．

図5.11(b-1)～(b-3) は従来の RIFCDP による整合度画像，図5.11(c-1)～(c-3) は重み減衰型 RIFCDP による整合度画像である．また，図5.11(b-1),(b-2),(b-3) では，表示階調を順に 0 ～ 1，0 ～0.125，二値化としている．これは，図5.11(c-1)～(c-3) でも同様である．二値化は検出率が最高となったしきい値 D_{thr} で行なっている．図5.11(b-1)，図5.11(c-1) の原画像では見分けにくかった右肩上がりの谷（黒い部分）線が，表示階調を 0 ～0.125 にすることで図5.11(b-2)，図5.11(c-2) のように見えてきている．

二手法の違いは，図5.11(b-3) と図5.11(c-3) を比較すると，より鮮明になる．この 2 つの図は，どちらも検出率が最高となったしきい値 D_{thr} で二値化しているものだが，RIFCDP では類似区間の一部しか残っておらず，逆に重み減衰型 RIFCDP では多くのノイズが存在する．これは，RIFCDP ではこの整合度の情報以外に検出区間の情報を持っているためであり，また重み減衰型 RIFCDP では，この段階では RIFCDP と異なり最小フレーム数 N_{min} の情報を用いていないためである．従って，重み減衰型 RIFCDP の類似区間検出処理では，この図5.11(c-3) の画像から右肩上がりで最小フレーム数 N_{min} 以上の長さの谷線を検出していることになる．

また，計算量を実測した結果，重み減衰型 RIFCDP では従来の RIFCDP の $\frac{1}{5.5}$ となり理論値の $\frac{1}{6.5}$ とほぼ一致した効果が得られた．また，このときのメモリ量は理論通り約 $\frac{1}{21}$ となった．これらの結果より，重み減衰型 RIFCDP ではメモリ量および計算量を削減しつつ，従来の RIFCDP と同様の検出率で類似区間を検出できることが示せた．

5.5 まとめ

局所距離にかかる重みを，過去に溯るに従って指数関数的に減少させて累積距離を計算することにより，計算量とメモリ量の軽減を可能とし，かつほぼ類似の機能をもつ 重み減衰型

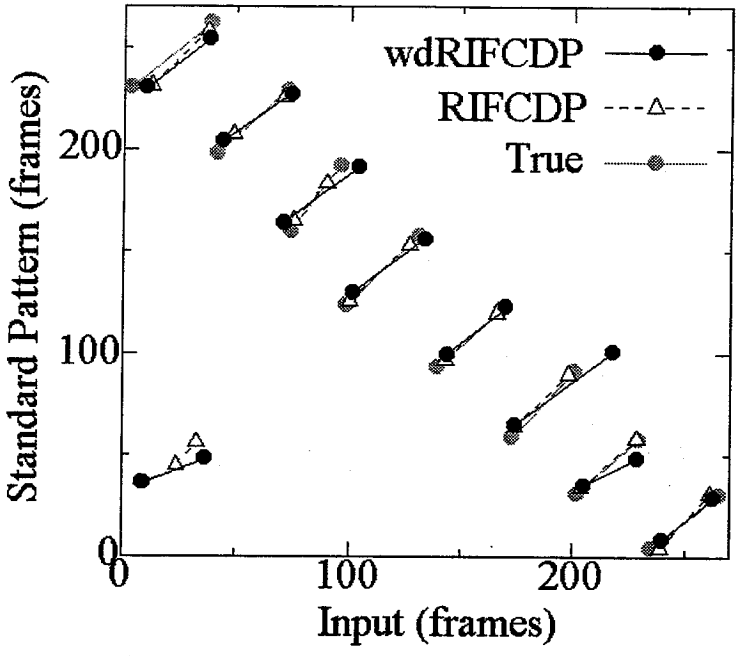


図 5.10 実験結果 [S1]

RIFCDP を提案した．また，この重み減衰型RIFCDPにより，従来のRIFCDPとほぼ同様な検出率を実現できることをジェスチャ動画像を用いた実験にて示した．このとき，計算量では $(\frac{1}{5.5})$，メモリ量では $(\frac{1}{21})$ とほぼ理論通りの効果を実証した．

今後は，手話動作をとらえた動画像データやマルチメディアデータに対して適用したいと考えている．

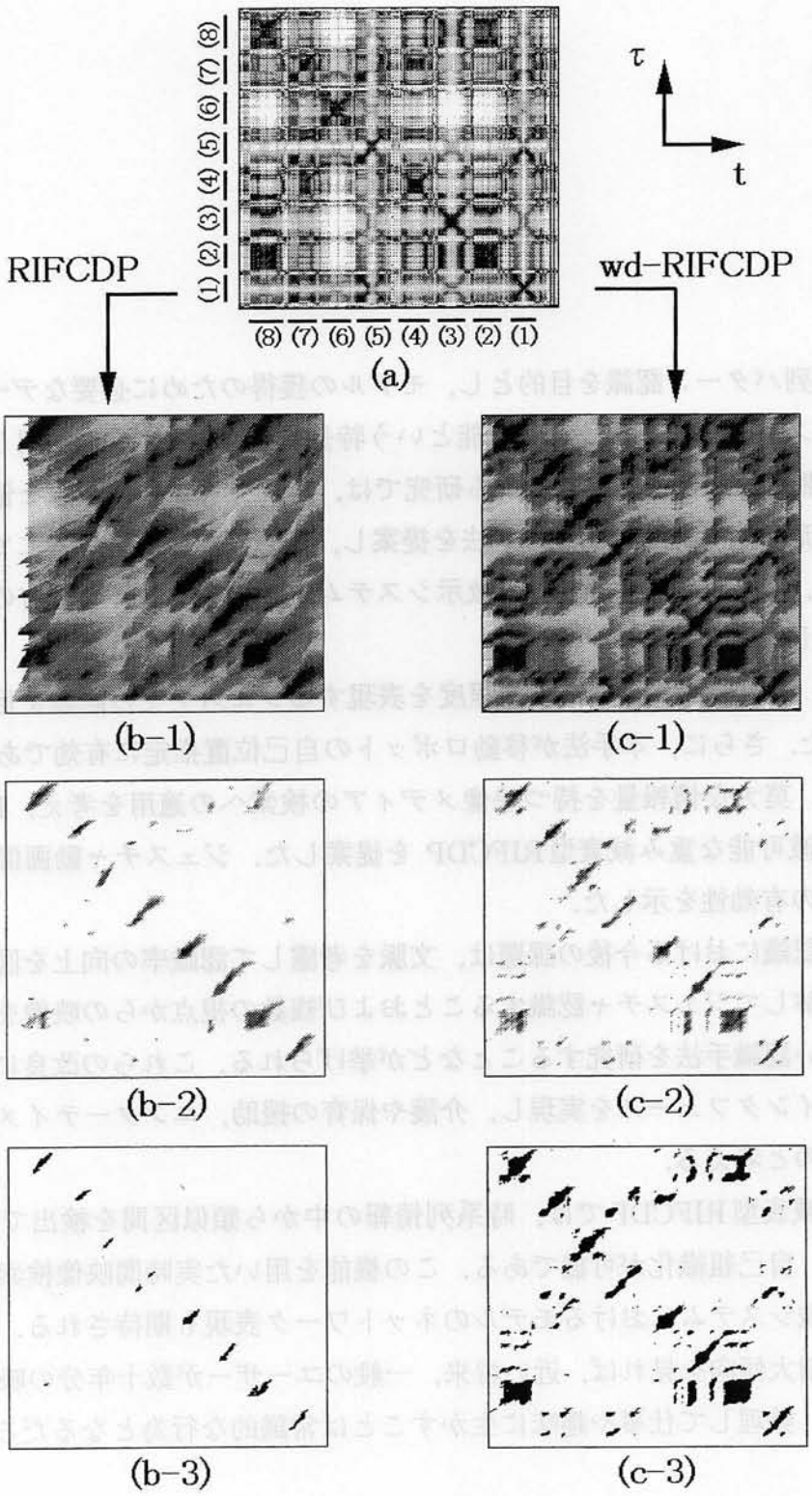


図 5.11 実験結果（画像）[S1]

第 6 章

緒言

動画像の時系列パターン認識を目的とし，モデルの獲得のために必要なデータが少なく，入力フレーム毎のスポッティング認識が可能という特長を持つ連続DPを活用した研究を行った．

第2章の人物のジェスチャを認識する研究では，動画像から必要十分な情報を取得するために手の動きと形状の特徴を抽出する手法を提案し，実時間認識システムにて本手法の有効性を示した．また，第3章にてオンライン教示システムを実現して動作者固有のジェスチャを容易に登録可能とした．

第4章では，戸惑ったジェスチャや程度を表現するジェスチャの認識を目指して非単調連続DPを提案した．さらに，本手法が移動ロボットの自己位置推定に有効であることを示した．

第5章では，莫大な情報量を持つ映像メディアの検索への適用を考え，RIFCDPの計算量，メモリ量を低減可能な重み減衰型RIFCDP を提案した．ジェスチャ動画像を用いた評価実験により本手法の有効性を示した．

ジェスチャ認識における今後の課題は，文脈を考慮して認識率の向上を図ることと移動している人物を追跡してジェスチャ認識することおよび複数の視点からの映像を用いて人物の向きに影響されない認識手法を研究することなどが挙げられる．これらの改良により，人にやさしいマンマシンインタフェースを実現し，介護や保育の援助，エンターテイメントの分野に活用されていくものと考える．

また，重み減衰型RIFCDPでは，時系列情報の中から類似区間を検出できるため，映像情報などの圧縮，自己組織化が可能である．この機能を用いた実時間映像検索システムの実現やジェスチャ認識システムにおけるモデルのネットワーク表現も期待される．特に，近年の計算機の記憶容量増大傾向を見れば，近い将来，一般のユーザーが数十年分の映像データから必要な情報を検索，整理して仕事や趣味に生かすことは常識的な行為となるだろう．

謝辞

本論文をまとめるまでに，多くの方々のご指導，ご協力を賜わった．

大阪大学白井良明教授には，1年以上にわたって貴重なお時間を頂き，多大なる御指導，御鞭撻を賜わった．ここに謹んで謝意を表します．特に筆者の行なった複数の研究間の関連性を見極め，新たな位置付けを与えて頂いた。また，浅田稔教授，岸野文郎教授，久野義徳助教授には，2回の審査を通して貴重な御指導を賜わったお陰により，本論文を高いレベルまで精錬することができた．ここに謹んで御礼申し上げます．

また，新情報処理開発機構つくば研究所情報統合研究室の岡隆一研究室長には，筆者の発表論文すべての作成において非常に貴重な助けを頂いた．島田潤一研究所長には，筆者の研究を絶えず暖かく援助していただいた．メディアドライブ(株)の野崎俊輔氏は，プログラム開発上最も重要な役割を果たして頂いた．つくば研のAndreas Held氏，向井理郎氏，伊藤慶明氏，小島浩氏，長屋茂喜氏，草柳明子氏，高橋信裕氏，森靖英氏，矢部博明氏，関本信博氏，櫻井茂明氏，伊原正典氏，およびメディアドライブ(株)の張健新氏，赤坂貴志氏，松村博氏との活発な議論が研究遂行の大きな原動力となった．篠崎啓助研究管理課長には，研究遂行上の心構えに関するご指導だけでなく論文冊子の作成においても貴重なお時間を頂いた．以上，つくば研究センタ，メディアドライブ(株)の諸氏に心から感謝の意を表する．

東京大学計数工学科藤村貞夫教授には，多くの時間を頂き筆者の研究に対する鋭いコメントを頂いた．安藤繁教授には，筆者が学部生として画像計測を始めるにあたってご指導を頂き，画像関係の研究を始める基礎を築くことができた．石川正俊教授，喜安千弥助手，元伊藤直史助手（現在群馬大学助教授)には，筆者が修士課程にて貴重な指導を頂いた．このように，大学の諸先生のお陰により，筆者が研究畑に踏み出すことがでたものであり，この場をお借りして深く御礼申し上げます．

日本鋼管(株)基盤技術研究所計測制御研究部の石野和成主査，壁矢和久主任研究員，押田栄二元主査，生澤勝美元部長，白井正明部長，元吉野正人副所長，長棟章生元主査上杉満昭主査には，多くの議論を通じて企業における研究と開発の方法についてご指導いただいた．現在，つくば研において研究を遂行できるのは，ひとえに日本鋼管(株)の諸氏のお陰であり，ここに心から感謝の意を表する．特に，石野主査には多大になご指導ご協力を頂いたにも関わらず，筆者からの貢献が少ないままに日本鋼管を退職する至ったことが申し訳ないと感じている．

奈良先端科学技術大学院大学助手山澤一誠氏，大阪大学谷内田正彦教授，八木康史助教授，

筑波大学大澤幸生助教授には，全方位視覚センサHyperOmni Visionの技術を快く提供して頂いただけでなく，研究上の心構えなど幅広いご指導を頂き，まことにありがたく感じております．また，小型全方位ミラーアタッチメントに関する最新の技術を快く提供して下さった京都大学石黒浩助教授に心から感謝いたします．

理化学研究所向井利春氏，大阪大学島田伸敬助手，九州大学内田誠一助手，中央大学梅田和昇助教授には，本論文の作成上有益な情報を頂き，また近い年代の研究仲間として交流していただき，深く感謝し致します．

参考文献

[1] D. H. Ballard, C. M. Brown, "Computer Vision", Prentice Hall(1982).

[2] 浅田　稔, "ダイナミックシーンの理解", 電子情報通信学会 (1994).

[3] 鳥脇　純一郎, "パターン認識と画像処理", 朝倉書店 (1992).

[4] 谷内田　正彦, "ロボットビジョン", 昭晃堂 (1990).

[5] 白井　良明, "連結領域を求める手法とそのハードウェア", 信学技法,IE78-9,1978.

[6] 大津 展之, 栗田 多喜夫, 関田 巌, "パターン認識", 朝倉書店 (1996).

[7] 宮武　孝文, 松島　整, 江尻　正員, "最大最小型画像フィルタリングの高速演算手法", 信学論 (D-II),Vol.J78-D-II,no.11,pp.1598-1607,1995.

[8] M. Ishikawa, "1 ms VLSI Vision Chip System and Its Applications", In Proc of FG'98,pp.214-215,Nara,Japan,Apr. 1998.

[9] 西村 拓一, 藤村 貞夫, 伊藤 直史, 喜安 千弥, "M配列を用いた鏡面物体の三次元形状計測", 電学論C, Vol.112-C, No.2, Feb., 1992 .

[10] 西村　拓一, 岡　隆一, "2次元連続DPによる画像のスポッティング認識", 信学技報,PRMU97-55,pp.1-7(1997-07).

[11] 西村　拓一, 岡　隆一, "動画像パターンへの重み減衰型時空間RIFCDP適用による移動物体の検出　- Self-applicative 連続DP -", 情報統合研究会,CII98-05,pp.1-7(1998-07).

[12] R.Bellman and R.Kalaba, "Dynamic Programming and Modern Control Theory", Academic Press, 1965.

[13] F. Jelinek, "Continuous speech recognition by statistical methods", Proc. IEEE, **64**, No.4, pp.532-556, 1976.

[14] S.E.Levinson, L.R.Rabiner and M.M.Sondhi, "An introduction to the application of the theory of probabilistic functions of a Markov process to automatic speech recognition", Bell Syst. J., **62**, 4, pp.1035-1074, 1983.

[15] 岡 隆一, "連続DPを用いた連続音声認識", 音響学会音声研資料, S78-20, pp.145-152(1978-06).

[16] 黒川　隆夫, "ノンバーバルインターフェース", オーム社(1994).

[17] 間瀬　健二, "顔とジェスチャの検出および認識", ロボット学会誌,Vol.16,no.6,pp.745-748,1998.

[18] V.I. Pavlovic, R.Sharma, T.S. Huang, "Visual Interpretation of Hand Gestures for Human-Computer Interraction:A Review," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 19, no. 7, pp.677-695, 1997.

[19] 石井 浩史, 望月 研二, 岸野 文郎, "人物像合成のためのステレオ画像からの動作認識法", 信学論(D-II),Vol.J76-D-II,no.8,pp.1805-1812,1993.

[20] 中嶋 正之, 柴 広有, "仮想現実世界構築のための指の動きの検出法", 信学論(D-II),Vol.J77-D-II,no.8,pp.1562-1570,1994.

[21] 亀田 能成, 美濃 導彦, 池田 克夫, "シルエット画像からの関節物体の姿勢推定法", 信学論(D-II),Vol.J79-D-II,no.1,pp.26-35,1996.

[22] 島田 伸敬, 白井 良明, 久野 義徳, "確立に基づく探索と照合を用いた画像からの手指の3次元姿勢推定", 信学論(D-II),Vol.J79-D-II,no.7,pp.1210-1217,1996.

[23] 岩井 儀雄, 八木 康史, 谷内田 正彦, "単眼動画像からの手の3次元運動と位置の推定", 信学論(D-II),Vol.J80-D-II,no.1,pp.44-55,1997.

[24] 近藤 拓也, 山際 貴志, 山中 光司, 山本 正信, "動画像からの動作感性情報の抽出", 信学論(D-II),Vol.J80-D-II,no.1,pp.247-255,1997.

[25] 渡辺 孝弘, 李 七雨, 谷内田 正彦, "インタラクティブシステム構築のための動画像からの実時間ジェスチャ認識-仮想指揮システムへの応用-", 信学論(D-II),Vol.J80-D-II,no.6,pp.1571-1580,1997.

[26] 牛田 博英, 山口 亨, 高木 友博, "ファジー連想記憶システムを用いた動作認識", 信学論(D-II),Vol.J77-D-II,no.8,pp.1562-1570,1994.

[27] J. Yamato, J. Ohya, K. Ishii, "Recognizing Human Action in Time-Sequential Images Using Hidden Markov Model", Proc. CVPR, pp.379-385, 1992.

[28] T. J. Darell and A. P. Pentland, "Space-Time Gestures", Proc.IJCAI'93 Looking at People Workshop(Aug. 1993).

[29] 高橋　勝彦, 関　進, 小島　浩, 岡　隆一, “ジェスチャー動画像のスポッティング認識”, 信学論 (D-II),Vol.J77-D-II,no.8,pp.1552-1561,1994.

[30] 大崎　喜彦, 山本　正信, “ステレオ画像からの3次元モデルのフィッティング”, 信学論 (D-II),Vol.J81-D-II,no.6,pp.1259-1268,1998.

[31] 西野　浩明, 凍田　和美, 宇都宮　孝一, “オンライン学習機能を備えた対話型両手ジェスチャインタフェース”, 信学論 (D-II),Vol.J81-D-II,no.5,pp.897-905,1998.

[32] 佐川　浩彦, 酒匂　裕, 大平　栄二, 崎山　朝子, 安部　正博, “圧縮連続DP照合を用いた手話認識方式”, 信学論 (D-II),Vol.J77-D-II,no.4,pp.753-763,1994.

[33] 内海　章, 大谷　淳, 中津　良平, “多数カメラを用いた手形状認識法とその仮想空間インタフェースへの応用”, 情処論,Vol.40,no.2,pp.585-593,1999.

[34] 林　健太郎, 久野　義徳, 白井　良明, “ユーザーの位置の拘束のないジェスチャによるヒューマンインタフェース”, 情処論,Vol.40,no.2,pp.556-566,1999.

[35] N. Shimada, Y. Shirai, Y. Kuno, and J. Miura, “Hand Gesture Estimation and Model Refinement using Monocular Camera - Ambiguity Limitation by Inequality Constraints”, In Proc of FG'98,pp.268-273,Nara,Japan,Apr. 1998.

[36] 呉　海元, 小林　弘知, 陳　健, 塩山　忠義, 島田　哲夫, “色彩動画像からの頭部ジェスチャ認識システム”, 情処論,Vol.40,no.2,pp.577-584,1999.

[37] 桐島　俊之, 佐藤　宏介, 千原　國宏, “プロトコル学習による身振りの実時間画像認識”, 信学論 (D-II),Vol.J81-D-II,no.5,pp.785-794,1998.

[38] 石淵　耕一, 岩崎　圭介, 竹村　治雄, 岸野 文郎, “画像処理を用いた実時間手振り推定とヒューマンインタフェースへの応用”, 信学論 (D-II),Vol.J79-D-II,no.7,pp.1218-1229,1993.

[39] 植田　健治, 石黒　浩, 辻　三郎, “パノラマ表現を用いた観測点の位置決めにおける一手法”, 信学論 (D-II),Vol.J75-D-II,no.11,pp.1809-1817,1992.

[40] 前田　武志, 石黒　浩, 辻　三郎, “全方位画像を用いた記憶に基づく未知環境の探索”, 情報処理学会研究会報告 CV-92-10,pp.73-80,1995.

[41] 石黒　浩, 山本　雅史, 辻　三郎, “能動的全方位視覚センサを用いた環境構造の復元”, 日本ロボット学会誌,9,5,pp.541-550,1991.

[42] 浅田　稔, “センサ情報の統合と理解による移動ロボットのための世界モデルの構築”, 日本ロボット学会誌,8,2,pp.28-38,1990.

[43] 速水 悟, 岡 隆一, "連続DPによる連続単語認識実験とその考察", 信学論(D),Vol.J67-D,no.6,pp.677-684,1984.

[44] 岡 隆一, "連続DPを用いた部分整合法によるフレーム特徴の音韻認識", 信学論(D),Vol.J70-D,no.5,pp.917-924,1987.

[45] 迫江 博昭, 藤井 浩美, 吉田 和永, 亘理 誠夫, "フレーム同期化, ビームサーチ, ベクトル量子化の統合によるDPマッチングの高速化", 信学論(D),Vol.J71-D,no.9,pp.1650-1659,1988.

[46] 中川 聖一, "拡張連続DP法による連続音声認識アルゴリズム", 信学論(D),Vol.J67-D,no.10,pp.1242-1249,1984.

[47] Y. Itoh, J. Kiyama, H. Kojima, S. Seki, R. Oka: "Reference Interval-free Continuous Dynamic Programming for Spotting Speech Waves by Arbitrary Parts of a Reference Sequence Pattern ", IEICE(D-II), J79-D-II, 9, pp.1474-1483(1996).

[48] J. Kiyama, Y. Itoh, R. Oka: "Topic-Independent Speech Summary and Automatic Topic Boundary Detection Using Incremental Reference Interval-free Continuous Dynamic Programming", IEICE(D-II), J79-D-II, 9, pp.1464-1473(1996).

[49] 古井 貞熙, "ディジタル音声処理", 東海大学出版会(1985).

[50] 中川 聖一, "確率モデルによる音声認識", 電子情報通信学会(1988).

[51] 柏野 邦夫, カビン スミス, 村瀬 洋, "ヒストグラム特徴を用いた音響信号の高速検索-時系列アクティブ検索法-," 信学論(D-II), Vol.J82-D-II,No.9,pp.1365-1373,1999.

[52] 西田,有木, "自動学習による話者セグメンテーション," 音声研資, SP97-57,pp.1-6,1997-11.

[53] 鷲尾, 緒方,有木, "ニュース音声に対するトッピングセグメンテーションの検討," 音学講論, 1-R-20,pp.157-158,1998-09.

[54] M. Sugiyama, "Retrieval of Acoustic Information," Technical Report of IEICE, SP99-25,1999-06.

[55] G. Smith, H. Murase, and K. Kashino, "Quick audio retrieval using active search," Proc. of ICASSP-98, Vol.6,pp.3777-3780,1998.

[56] Y. Itoh, J. Kiyama, and R. Oka, "Speech understanding and Speech retrieval for TV program based on spotting algorithms," Proc. of Spring meeting of ASJ,3-P-22(1995-03)[In Japanese].

[57] Y. Itoh, J. Toyoura, J. Kiyama, and R. Oka, "Real-time retrieval from speech or text database with spontaneous speech by using Reference Interval-free Continuous DP," Proc. of Autumn meeting of ASJ,3-P-22(1995-09)[In Japanese].

[58] 中村裕一，外村佳伸，"見たい部分を簡単に短時間で-気の利いた映像メディア技術を目指して-"，電子情報通信学会誌, vol.82, No.4, pp.346-253, 1999.

[59] 兵後裕子，中川聖一，"連続発生された二発話文間のDPマッチングによる共通部分の抽出", 1989信学春季全大，A-22,March 1989.

[60] 伊藤慶明，木山次郎，小島浩，関進，岡隆一，"時系列標準パターンの任意区間によるスポッティングのための Reference Interval-free 連続DP(RIFCDP)", 信学論(D-II), J79-D-II, 9, pp.1474-1483(1996).

[61] 木山次郎，伊藤慶明，岡隆一，"Incremental Reference Interval-free 連続DP による任意話題音声の要約と話題境界検出", 信学論(D-II), J79-D-II, 9, pp.1464-1473(1996).

[62] 伊藤　慶明，木山　次郎，関　進，小島　浩，張健新，岡　隆一，"同時複数話者の会話音声およびジェスチャのリアルタイム統合理解による Novel Interface System", 音声言語情報処理,7-3,pp.17-22,1995.

[63] J.A.Mclaughlin and J.Raviv, "Nth-order autocorrelations in pattern recognition," Information and Control, Vol.12, pp.121-142, 1968.

[64] J.L. Crowley, "World modeling and position estimation for a mobile robot using ultrasonic raging," In Proc. IEEE Int. Conf. on Robotics and Automation, pp.674-680, 1989.

[65] H. Takeda, C. Facchinetti, and J.C.Latombe, "Planning the motions of a mobile robot in a sesory uncertainty field," IEEE Trans. Pattern Anal. Machine Intell., 16(10),pp.1002-1017,1994.

[66] H. Xu, H. van Brussel, J. de Schutter, and J. Vandrope, "Sensor fusion and positioning of the mobile robot LiAS," In U.Rembold et al. editor, Intelligent Autonomous Systems, pp.246-253, IOS Press, 1995.

[67] C. Facchinetti, F. Tieche, and H. Hugli, "Self-positioning robot navigation using ceiling image sequences," In Proc. 2nd Asian Conference on Computor Vision, Vol.III, pp.814-818,Singapore,1995.

[68] T. Maeda, H. Ishiguro and S. Tuji, "Memory-Based Navigation using Omni-directional View in Unknown Environment," IPSJ ,CV-92,pp.73-80,1995.

[69] Y. Matsumoto, M. Inaba and H. Inoue, "Navigation of A Mobile Robot for Indoor Environments Basen on Scene Image Sequence," MSJ Convention meeting , Vol.A,pp.481-484,1995.

[70] 山澤　一誠, 八木　康史, 谷内田　正彦, "移動ロボットのナビゲーションのための全方位視覚センサHyperOmni Visionの提案," 信学論(D-II),Vol.J79-D-II,no.5,pp.698-707,1996.

[71] R.Sharma, T.S. Huang, V.I. Pavlovic, Y. Zhao, "Speech/Gesture Interface to a Visual Computing Environment for Molecular Biologists," Proc. Int'l Conf. Pattern Recognition,1996.

[72] J. Y. Zheng and S. Tsuji, "Panoramic Representation for Route Recognition by a Mobile robot," International Journal of Computor Vision,9:1,pp.55-76,1992.

[73] S.W.Bang, W.Yu, and M. J. Chung, "Sensor-Based Local Homing Using Omni-directional Range and Intensity Sensing System for Indoor Mobile Robot Navigation," Preceedings of the 1995 IEEE/RSJ International Conference on Intelligent Robots and Systems, Vol.2, pp.542-548,1995.

[74] Andreas Held and Ryuichi Oka, "Sonar Based Map Acquisition and Exploration in an Unknown Office Environment," International Symposium on Intelligent Robotic Systems 95, Nov. 1995.

[75] 小島浩, 伊藤慶明, 岡隆一, "Reference Interval-free 連続DPを利用した移動ロボットの時系列画像による位置同定システム," 信学論(D-II), J80-D-II, 3, pp.724-733(1997).

発表論文

第2章

[1] 西村 拓一, 向井 理朗, 野崎 俊輔, 岡　隆一, "低解像度特徴を用いた複数人物によるジェスチャの単一動画像からのスポッティング認識", 信学論 (D-II),Vol.J80-D-II, No.6, pp.1563-1570,1997.

[2] T. Nishimura and R. Oka, "Spotting Recognition of Human Gestures from Time-Varying Images", International Conf. on Automatic Face and Gesture Recognition, pp.318-322(1996-10).

[3] T. Nishimura and R. Oka, "Towards the Integration of Spontaneous Speech and Gesture based on Spotting Method", International Conf. on Multisensor Fusion and Integration for Intelligent Systems, pp.433-437(1996).

[4] T. Nishimura ,T. Mukai, R. Oka, "Spotting Recognition of Human Gestures performed by People from a Single Time-Varying Image", 7th International Conf. on Human-Computer Interaction, pp.33(1997-8).

[5] T. Nishimura ,T. Mukai, R. Oka, "Spotting Recognition of Human Gestures performed by People from a Single Time-Varying Image", IROS'97, vol. 2, pp.967-972(1997-9).

[6] 西村 拓一, 向井　理朗, 岡　隆一, "白黒動画像からの形状特徴を用いたジェスチャのスポッティング認識システム," 信学論 (D-II), Vol.J81-D-II,No.8,pp.1812-1821,1998.

第3章

[7] 西村 拓一, 向井 理朗, 野崎 俊輔, 岡　隆一, "動作者適応のためのオンライン教示可能なジェスチャ動画像のスポッティング認識システム," 信学論 (D-II), Vol.J81-D-II,No.8,pp.1822-1830,1998.

[8] T. Nishimura, Hiroaki Yabe , and R. Oka, "A Method of Model Improvement for Spotting Recognition of Gestures Using an Image Sequence," New Generation Computing, to be published.

第4章

[9] 西村 拓一, 野崎　俊輔, 向井　理朗, 岡　隆一, "連続DPへの非単調性導入による ジェスチャ動画像からの戸惑い動作のスポッティング認識," 信学論 (D-II),Vol.J81-D-II, No.1, pp.18-

26,1998.

[10] T. Nishimura , T. Mukai , R. Oka, "Non-monotonic Continuous Dynamic Programming for Spotting Recognition of Hesitated Gestures from Time-Varying Images", ACCV'98,vol.II,pp.734-741(1998-1).

[11] 西村 拓一, 野崎　俊輔, 岡　隆一, "Non-monotonic 連続 DP によるスポッティングに基づく移動ロボットの時系列画像を用いた大局的位置の推定," 信学論 (D-II), Vol.J81-D-II,No.8,pp.1876-1884,1998.

[12] T. Nishimura, H. Kojima, Y. Itoh, A. Held, S. Nozaki, S. Nagaya, and R. Oka, "Effect of Time-spatial Size of Motion Image for Localization by using the Spotting Method," IEEE, Proceedings of ICPR'96, pp.191-195,1996.

第5章

[13] 西村 拓一, 古川 清, 向井 理朗, 岡　隆一, "時系列パターン検索のための重み減衰型 RIFCDP について", 信学論 (D-II),Vol.J81-D-II, No.3, pp.472-482,1998.

[14] Takuichi NISHIMURA, Kiyoshi FURUKAWA, Toshiro MUKAI,Ryuichi OKA, "Weight-decreasing Reference Interval-Free Continuous DP for Retrieval of Time-Sequence Pattern", Systems and Computers in Japan, Vol.29, No.10, pp.15-25,1998.