NAIST-IS-DT9861007

博士論文

注視点の学習と選択制御に基づく身振りの実時間画像認識

桐島 俊之

2000年2月7日

奈良先端科学技術大学院大学
情報科学研究科 情報処理学専攻

本論文は奈良先端科学技術大学院大学情報科学研究科に
博士 (工学) 授与の要件として提出した博士論文である.

桐島 俊之

審査委員： 千原 國宏 教授
植村 俊亮 教授
横矢 直和 教授
竹村 治雄 助教授
眞鍋 佳嗣 助教授

注視点の学習と選択制御に基づく身振りの実時間画像認識*

桐島 俊之

内容梗概

本研究の目的は，非接触・非装着型身振りインタフェースの実現に不可欠である任意人物の任意身振りを画像により実時間で認識する手法を提案することにある．具体的には，身振り画像の特徴選択とサンプリングの問題を解決するための手法を提案する．本論文は以下の６章により構成される．

第１章では，身振り認識技術が必要とされる背景に言及した後，従来の身振り計測技術と画像認識手法を取り上げ，その問題点を指摘し，本研究における課題を明らかにする．

第２章では，認識の対象となる身振りの種類や性質を限定せず，広範な身振りの認識を可能とするために，多注視点身振り認識法を提案する．提案手法により，衣服や身振りの種類のみならず軌跡の大小・動作特性の個人間変動に対して頑健な認識処理が実現されることを，評価実験により実証する．

第３章では，多注視点身振り認識法において対応の困難な認識処理速度の低下および不安定化の問題を解決するために，多注視点選択制御法を提案する．ジェスチャービデオシステムを利用した評価実験により，常時 30[frames/s] でのインタラクションが実現されることを示し，認識処理速度の問題について提案手法が有効であることを実証する．

第４章では，手話画像データベース検索システムの構成方法について述べ，提案手法による画像検索機能の実現可能性を示す．提案システムを単一プロセッサ

*奈良先端科学技術大学院大学 情報科学研究科 情報処理学専攻 博士論文，NAIST-IS-DT9861007, 2000年2月7日．

のパーソナルコンピュータ上に実装し，864 枚の手話動作画像から構成される画像データベースについて最短 4.2[s] での高速な検索処理が実現されることを示す．

第 5 章では，提案手法の長所と短所について考察と検討を加えた後，画像認識による身振りインタフェースの実用化に向けて今後解決されるべき課題に言及する．

第 6 章では，本研究で明らかにした事項について総括する．

キーワード

身振り認識，身振りインタフェース，身振りプロトコル，多注視点身振り認識法，多注視点選択制御法

Realtime Gesture Recognition by Learning and Selective Control of Visual Interest Points*

Toshiyuki Kirishima

Abstract

The prime objective of this research is to establish a framework to recognize arbitrary person's unspecified gestures in realtime, which realizes a vision-based gesture interface for virtual reality. This dissertation presents a solution to the problems of (1) feature selection (2) sampling rate control of gestural image sequence and consists of following six chapters.

In chapter 1, the background and the need for gesture recognition techniques are briefly mentioned. After the introduction of typical examples of conventional study, the neglected problems are elucidated in detail, which are the target issues in this research.

In chapter 2, a recognition framework called QVIPS is proposed, which does not assume the types and natures of gestures to be recognized. Experimental results demonstrate that QVIPS achieves robust recognition against such factors as types of clothes and gestures, the extent of motion trajectories and individual differences of motion characteristics.

As QVIPS devoids of self-load monitoring and controlling functionality, a selective control method for visual interest points is proposed in chapter 3, which deals with the problem of slow and unstable processing speed of the recognition system. A gesture video system is developed to show that the proposed methods

*Doctor's Thesis, Department of Information Processing, Graduate School of Information Science, Nara Institute of Science and Technology, NAIST-IS-DT9861007, February 7, 2000.

can overcome the aforementioned problem and achieve full video rate interaction with displayed objects.

To suggest feasibility for a retrieval task by the methods presented above, a retrieval system for sign language image database is proposed in chapter 4. The proposed system is implemented on a personal computer with single processing unit and accomplishes the shortest search time of 4.2[s] for 864 images in the sign database without any special-purpose hardware.

In chapter 5, detailed considerations are given to the merits and demerits of the proposed methods. Subsequently, some of the unavoidable problems in putting vision-based gesture interface to practical use will be pointed out.

Chapter 6 briefly concludes the presented ideas and achievements in this work.

Keywords:

Gesture recognition, Gesture interface, Gesture protocol, QVIPS, Selective control of visual interest points

目 次

図 目 次

表 目 次

1 序論

1.1 身振り認識技術への期待

近年の著しいコンピュータ技術の進歩と普及は，仮想現実感技術による新たな情報環境の構築を可能とし，医療・福祉・教育・娯楽・建築・設計など，我々の生活を身近に支える分野において大きな革新をもたらしつつある．特に，インターネット技術の普及に伴い，情報ネットワーク上のサイバースペースは急速な拡大の一途にある．

サイバースペースにおける利用者インタフェースは，そこで提供される情報サービスの質を左右する重要な要因の一つである．例えば，商品購入時のブラウジング方式において，静止画像のみによるものと，商品を意のままに操ることができる仮想現実感技術によるものとでは，顧客を説得する度合いのみならず購入時の満足度さえも左右することは容易に推測できる．また，電子図書館における資料の閲覧や，仮想博物館・美術館における電子展示物の鑑賞の際に，我々の日常的な振る舞いをそのまま利用できれば，蓄積された膨大な情報へのアクセスは一層容易になると同時に，こうした経験を我々は情報ネットワークを媒介して共有することが可能となる．仮想現実感技術は，我々の日常生活における情報獲得の形態そのものを変革する可能性を持っており，インターネット上のサイバースペースにおいても必須の技術となることは間違いない．

上述の例からも明らかであるように，仮想現実感技術は，これまで情報共有の場であったサイバースペースを，経験共有の場に変貌させる原動力となる可能性をも秘めている．従って，専門家のみならず年少者から高齢者までの広範かつ多様な人々に仮想現実感技術が利用される状況をあらかじめ想定し，我々が日常的に利用している音声・顔表情・身振りなどのメディアチャネルに対応する，より

柔軟で高度なヒューマンインタフェースを実現する必要がある [1, 2]．仮想現実感環境におけるインタラクションのリアリティを高め，利用者に仮想現実感環境への知覚的な没入感を与えるためにも，利用者とコンピュータとの間を媒介するインタフェース技術は重要である．

顔表情や音声を除いた利用者の非言語的な情報は，大半が身体動作により表出されるため，サイバースペースにおける利用者インタフェースは，時々刻々と変化する利用者の身体動作を実時間で認識する必要がある．身振りを媒介とした非言語的コミュニケーションは，原始的である反面，豊かな情報量を持ち，人間にとって言語と並んで重要なものである [3]．従って，仮想現実感環境における人間－機械系インタラクションにこうした非言語的な情報を活用するのは必然である．

従来，顔表情 [4] や音声をコンピュータに認識させるための研究は数多くなされ，実用的な商用システムも開発されている一方で，利用者の身体動作をコンピュータに認識させるための研究は，仮想現実感技術との関連からその重要性が近年になって広く認知されたに過ぎず，まだその途に就いたばかりである．

1.2 身振り計測技術

従来から身振りの計測には用途や目的に応じた装着型インタフェースが用いられている．手指の微妙な曲げなどを検出するためにグローブ型センサが利用され，頭部や腕の関節などの空間的位置を検出するために電磁気式の点位置検出センサが利用されている．これらの身体位置・姿勢計測装置は，一般に装着型身振りインタフェースと呼ばれる [5, 6, 7, 8, 9]．

装着型身振りインタフェースは，限定的な身体部位の位置や姿勢の検出を可能にするが，利用者を仮想現実感システムへと物理的に拘束してしまう問題点がある．この問題点の他にも，

［１］計測範囲が狭く装脱着が必要である上に，事前の煩雑なキャリブレーション作業が必要であること，

［２］グローブ型センサや電磁気式センサを併用する必要があり，それぞれのセンサ特有の問題点を回避できないこと，

［３］全身の位置や姿勢を計測する必要がある場合，多数のセンサの装着が必要となるため，利用者側の身体的負担が増大すること，

［４］不特定多数の利用者が使用する場面，あるいは複数の利用者が同時に使用する場面を想定すると，装置自体の購入コストや装置使用における衛生面での問題点を回避することが出来ないこと，

などの多数の問題点がある．

身振りを認識する際に必要な情報は，認識対象の動作の性質に強く依存することは明らかであり，不要な情報は後で無視すれば良いのであるから，身振りを計測する際には，高次処理の段階で役立つ可能性のある情報まで幅広く計測しておく必要があると考えるのが妥当であろう．このことからも，身振り計測にはＣＣＤカメラなどの画像入力手段を利用するのが望ましいと言える．

画像処理に基づく身振りインタフェースの実現により，上記問題点の解決が期待できる．画像処理に基づく身振りインタフェースは，身振りを撮影した画像（以降，身振り画像と呼ぶ）から利用者の非言語情報を認識するが，画像から人物の姿勢や動きを認識するには様々な課題を克服する必要があり，現状では研究の途上にある．

1.3 身振り画像認識手法

本研究では，非接触・非装着型身振りインタフェースが次世代ヒューマンインタフェースとして重要な役割を担うと考え，画像処理に基づき身振りを認識するインタフェースの実現手法に焦点を当てる．本節では画像処理に基づく身振り認識手法を表 1.1 に示すように，身体モデルに基づく手法と図形パターンに基づく手法に大別して紹介する．

身体モデルに基づく手法では，実画像中の人物姿勢を推定することを主目的としており，対象物体とコンピュータ内部の幾何学的モデルとの対応付けを求めることが重視される．

一方，図形パターンに基づく手法の目的は，実画像中の人物像のパターンや領域情報に基づいて，身振り画像のクラスをパターン認識手法により推定することにある．表 1.1 に示した代表的手法について，以降の節で個別に紹介する．

表 1.1 身振り画像認識の代表的手法

分類	代表的手法
身体モデルに基づく手法	幾何学的モデルに基づく手法
	色情報とパターンの位置関係に基づく手法
	遺伝的アルゴリズムに基づく手法
図形パターンに基づく手法	連続ＤＰ法に基づく手法
	ファジー連想記憶に基づく手法
	隠れマルコフモデルに基づく手法
	ＫＬ展開に基づく手法
	図形モーメントに基づく手法

1.3.1 身体モデルに基づく手法

（１）幾何学的モデルに基づく姿勢推定手法

山本らは，人体ＣＡＤモデルを実画像へとフィッティングさせることにより，人物像の追跡とその姿勢を推定する手法を提案している [10, 11]．提案手法では，追跡開始フレームにおいて人体ＣＡＤモデルを人手により実画像内の人物像に一致させ，その後，運動パラメータの推定とモデルの移動を交互に行うことにより，人物動作を追跡している．

問題点としては，モデルと人物像との不一致が発生する度に誤差が蓄積されるため，最終的には追跡に失敗することが挙げられる．誤差を生じさせる主な要因としては，

［１］人体モデル自体が不完全であること，

［２］追跡時にモデルの進み過ぎや遅れによる背景物体像の混入があること，

の２点が挙げられる．

この問題に対処するために，実画像とＣＡＤモデルとの照合を幾つかのキーフレームにおいて行い，その結果得られた人物像の位置と姿勢の拘束を中間フレームに伝搬させる手法が提案されている [12]．提案手法では，円盤投げの投てき動

作の追跡に成功している．しかしながら，キーフレームにおけるモデル照合については人手で行わなければならない問題がある．

なお，上記手法以外にも，大垣らによる形状モデルと動きモデルを利用した時系列ステレオ画像からの人物姿勢推定手法 [13]，米元らによる多視点画像を利用した人物姿勢の追跡手法 [14] が提案されている．

（２）色情報とパターンの位置関係に基づく姿勢推定手法

石淵らは，色情報を利用して入力画像から手領域を切り出し，領域内の２値画像に基づいて主軸検出・回転変換・手領域重心検出・指先検出を行うことにより，手の位置や姿勢パラメータを推定する手法を提案している [15]．

実時間での手振り推定を実現するために，パイプライン方式による並列処理を行い，オンラインで動作する認識システムを構築している．入力画像の大きさは横 512[dot] 縦 480[dot] であり，並列化によりシステム全体で平均 10[Hz]（8 〜 12[Hz]）の動作周波数を達成しているが，手振り動作への応答に 200[ms] 程度の遅延が生じる問題がある．

（３）遺伝的アルゴリズムに基づく姿勢推定手法

大谷らは，遺伝的アルゴリズムに基づいて，実画像中の人物の上半身姿勢を推定する手法を提案している [16]．この研究では姿勢パラメータ推定を，組み合わせ最適化問題と位置付けて遺伝的アルゴリズムを適用している．提案手法では，染色体内の遺伝子の値に従って３次元の人体モデルを変形・合成し，実際に撮影された人物の多視点画像との整合の度合いを評価している．

提案手法の場合，身体動作に関する制約や知識が不要であり，さらに，動作画像列がなくても姿勢の推定を行えるメリットがある．ただし，腕などの頻繁に隠蔽が発生する身体部位に対して正しく推定されない場合がある．

（４）まとめ

身体モデルに基づく手法は，撮影条件や照明条件などを厳密に考慮する必要がある上に，モデルフィッティングの際に煩雑な収束計算を必要とすることが多い．また，人物の着衣に制限が加えられたり，追跡の開始前に初期パラメータを人為

的に設定する必要があることが多い．さらに，撮影角度や認識対象の動作によっては頻繁に隠蔽が発生し，身体モデルパラメータの推定処理自体が不安定となる問題がある．

身体モデルのパラメータ推定には通常数百ミリ秒から数十秒を要し，推定精度はモデルフィッティング手法の成否に依存する．また，推定パラメータの収束時間を予測することが困難であり，一定の時間内で認識結果を出力することが要求される仮想現実感システムのための身振りインタフェースには，現状では適していない．

1.3.2 図形パターンに基づく手法

（１）連続ＤＰ法に基づく手法

西村らは，リダクション画像と呼ばれる特徴画像に基づいて身振りを認識させる手法を提案している [17, 18, 19, 20, 21]．音声認識の分野でよく利用される連続ＤＰ法が，標準画像列と入力画像列とを照合する際に適用される．

この研究では，動作の違いが明確な８種類の独自定義の身振り動作について，個人差や衣服に依存しない認識処理を実現している．提案手法は，ワークステーション上に実装され，入力画像の大きさ横 160[dot] 縦 120[dot] において 15[frames/s]（以降，[frames/s] は [fps] と表記）程度での認識処理を実現している．動作の種類が認識結果として出力され，これはジェスチャコマンドとして利用される．

問題点としては，例えば縦 12[dot] 横 12[dot] のように入力画像を極端に低解像度変換してしまうため，細やかな動作の認識に対応できないことが挙げられる．また，同時に考慮される特徴画像が１種類のみであるため，認識可能な身振りの性質や種類があらかじめ制限される問題もある．

（２）ファジー連想記憶に基づく手法

牛田らは，認識対象の動作に対応するメンバーシップ関数とファジー推論ルールを連想記憶システムにあらかじめ埋め込むことにより，身振り動作を認識する手法を提案している [22]．不特定人物の３種類のテニス動作について，従来のファジー推論手法やニューラルネットワーク手法に比べて良好な認識結果が得られる

ことを示している．

テニス動作については，色情報に基づいて右肩と右手先を検出した後，それぞれの重心座標を求め，胴体と腕の成す角度 θ を推定し，これを特徴量として用いている．また，５種類の指示動作の認識実験を行い，提案手法が不特定者動作の認識に適していることを示している．

提案手法では，不特定者の動作認識のためにあらかじめ個人に依存しない特徴量の選定がなされ，その特徴量の時系列波形のピーク値が連想記憶システムへと入力されている．問題点としては，時系列波形のピーク値のみしか考慮していないため，動作軌跡の大小に頑健な認識を行うことが困難であることが挙げられる．また，認識対象の動作に特化した単一特徴量の選定とファジー推論ルールの事前作成が不可欠であるため，認識対象の動作の種類や性質が既知でない場合，提案手法を適用することが出来ない問題がある．

なお，上記手法以外にも，小俣らによるファジー推論を利用した手振り認識手法 [23] が提案されている．

（３）隠れマルコフモデルに基づく手法

大和らは，隠れマルコフモデル（以降，HMMと呼ぶ）を利用してテニス動作を認識する手法を提案している [24]．学習には離散HMMが用いられ，特徴ベクトルとしては人物領域のメッシュ特徴を使用し，これをベクトル量子化（以降，ＶＱと呼ぶ）して得られるシンボルの時系列をHMMに入力することにより特徴的動作を分類している．

６種類のテニス動作に対して 90(%) 程度の認識率が得られているが，不特定者の動作については認識率が低いという問題がある．そこで，各カテゴリーのHMMごとにＶＱコードブックを用意し，シンボル列の生成をカテゴリー別に行うことで，上記問題点の解決が試みられている [25]．その結果，全カテゴリーで共通のＶＱコードブックを持つ離散HMMを用いる場合よりも，認識率が向上することが確認されている．入力画像は縦 200[dot] 横 200[dot] の濃淡画像であり，この画像は後に縦 8[dot] 横 8[dot] の特徴ベクトルに変換された後，オフラインで認識処理される．

なお，上記手法以外にも，畠らによるＫＬ展開とHMMを組み合わせた身振り

認識手法 [26]，Pentland によるマルチモーダルシステムのためのHMMを用いた人物動作認識手法 [1]，Starner らによるHMMを用いたＡＳＬ（American Sign Language）手話文字の認識手法 [27] が提案されている．

（４）ＫＬ展開に基づく手法

藤本らは，ＫＬ展開に基づいて画像系列を固有空間に射影し，固有空間内での軌跡の類似性を評価することにより，例えば身振りのような複数枚の画像により表現される対象を検索する手法を提案している [28]．提案手法では，ＣＬＡＦＩＣ法に基づいて入力画像と動画像データベース内の各固有空間の距離を求め，類似系列における軌跡の類似性を評価している．同一人物の８種類の独自定義かつ単色背景の身振り画像について良好な検索結果が得られている．

複数人物動作の検索や衣服に依存しない検索が必要な場合，固有空間の次元数を経験的に設定する必要がある．また，ＫＬ展開を入力画像に直接適用するため，動作位置のずれや動作特性の変化，背景ノイズに対する頑健さを期待することは困難である [29, 26]．なお，処理対象画像は，横 360[dot] 縦 238[dot] の濃淡画像であり，検索はオフライン実行されている．

上記手法以外にも，渡辺らによるテンプレートマッチング法とＫＬ展開を組み合わせた身振り認識手法 [30]，Cui らによるＭＤＦ特徴などの判別分析特徴に基づいたＡＳＬ手話文字を認識する手法 [31] が提案されている．

（５）図形モーメント特徴に基づく手法

E.Hunter らは，Zernike Moments（以降ＺＭと呼ぶ）と呼ばれる図形モーメント特徴による手振り認識手法を提案している [32]．以下にＺＭの定義を示す．

$$Z_{nl} = \frac{n+1}{\pi} \int_0^{2\pi} \int_0^1 V_{nl}(r,\theta) f(r\cos\theta, r\sin\theta) r dr d\theta \tag{1.1}$$

ここで n は次数，l は繰り返し数である．また，ＺＭ多項式は，

$$V_{nl} = V_{nl}(r\cos\theta, r\sin\theta) = R_{nl}(r)\exp(jl\theta) \tag{1.2}$$

により定義される．ＺＭにより回転不変の形状特徴量が得られるが，ＺＭはその次数 n と繰り返し数 l 次第で個人に依存する形状特徴量にもなり得るし，個人に

依存しない形状特徴量にもなり得るため，あらかじめ訓練サンプルについて判別分析するなどして最適な次数 n と繰り返し数 l を決めておく必要がある．なお，ＺＭを画像解析や認識に応用する際の問題点やその限界については文献 [33] に詳しく述べられている．

この研究では，同一人物の６種類の単色背景の手振り画像を訓練サンプルとして，個人差に比較的影響を受けない認識処理を実現している．なお，提案手法は，ワークステーション上に実装され，オンラインで 1[fps] 程度の処理速度を達成している．

（６）まとめ

図形パターンに基づく手法では，事前に認識対象身振りの標準パターン登録などの手続きが必要であるが，画像１枚につき数十ミリ秒から数秒で処理されるため，身振り画像の実時間認識処理に適している．この理由としては，撮影条件や照明条件などを厳密に考慮する必要がないことと，幾何学的モデルとのフィッティング処理が不要であることが挙げられる．また，利用者に特別な目印や装置の着用を求める必要がなく，認識処理が隠蔽の影響を受けることもない．

利用者の身振り動作に仮想現実感システムが即座に応答することは，インタラクションのリアリティを高める上で非常に重要である [34, 35] ため，本研究ではこの方式による身振りインタフェースの実現を目指したい．しかしながら，専用ハードウェアを導入せずにソフトウェア処理のみでビデオレート（30[fps]）での認識処理を実現する研究例は少ない [36] のが現状である．

1.4 本研究における課題

1.4.1 身振り画像の特徴選択問題

従来手法では認識処理の高速化を図るために，目的の達成に最適と思われる単一の特徴画像やモーメント特徴，さらに関節角などの特徴を，事前にシステム開発者が検討した後で，認識手法が開発され実装されている場合が多い．特徴画像の例としては，リダクション画像，シルエット画像，差分画像，エッジ・輪郭画像，オプティカルフロー画像，レンジファインダなどによる距離画像 [37]，赤外線カメラによる赤外線画像などが挙げられる．特徴画像は，画像の取得原理や特徴抽出方法に依存して実際には無数に生成できるため，上述の例は特徴画像の代表例に過ぎない．従来手法では単一の特徴量しか同時に考慮していない場合が多く，このため身振り動作の再現性がシステムの認識性能に直接悪影響を及ぼしてしまう問題がある．また，従来手法では認識の対象となる身振りの種類や性質をあらかじめ限定する場合が多く，広範な身振りの認識には対応できないことも問題である．

普段我々は，重要な特徴を選択的に注視することにより，標準的な身振りのみならず，それと類似した身振り（以降，類似身振りと呼ぶ）の認識さえも労なく行っている．類似身振りとは，標準的な身振りの特徴を多く含む身振りであり，我々が日常的場面で行使している身振りのほとんどは類似身振りであると言っても過言ではない．

類似身振りの認識のために，標準的な身振りに類似するすべての動作を逐次追加記憶させる方式がまず初めに考えられる．当然，この方式では記憶すべき画像枚数とそれに対応するパターン照合コストが膨大となるため，実時間での認識処理が困難になるという問題点がある．

一般に，同一の身振りとして定義可能な動作群には，それらの視覚的特徴において何らかの共通項が存在することが多い．もし，重要な共通項を注視することができれば，類似身振りをすべて学習する必要がなくなるため，パターン照合のコストを一定に保つことができる．さらに，学習時に示さなかった類似身振りに対する認識率の向上が期待できるのみならず，無用な特徴を無視することにより認識処理の高速化が可能となるメリットもある．この問題は，心理学の立場から

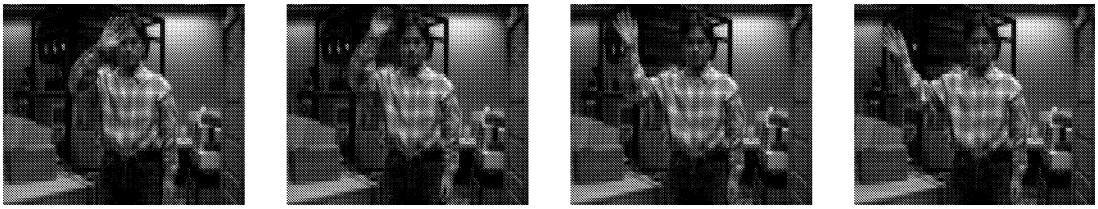
焦点選択の問題として古くから研究され，その重要性と難しさは認識されているが，具体的に工学的立場からこの問題を扱っている研究例はまだまだ少なく，特に身振り認識に関してはほとんど見当たらない．この問題を具体的に図 1.1 を用いて説明する．

図 1.1 の各写真はすべて正面を向いて片手を振っている人物のスナップショットである．もし，図中の身振り動作をすべて同一とみなすならば，この動作には例えば「バイバイ（さようならの意）」といった意味付けが可能である．この場合「手を振る」という時間差分特徴に対応する視覚的特徴を注視すれば良い．しかし，図中の身振り動作を別々の動作とみなすならばこうした意味付けがなされることはない．この場合は人物の位置を注視して，各動作の意味付けを行えば良いことになる．

上記の例からも明らかであるように，身振り動作とその意味付けは，人間の意向を無視しては一意に確定しないため，人間の意向を反映させるための学習が不可欠である．こうした学習と認識を実現するには，同一のクラスに属する身振りであるとして与えられた身振り画像から，時空間的に共通または安定した特徴群を見出し，その特徴を注視する選択的な身振り認識処理の枠組みが不可欠となる．特定の特徴を注視するためには，身振り認識における注視機構の問題に取り組む必要がある．

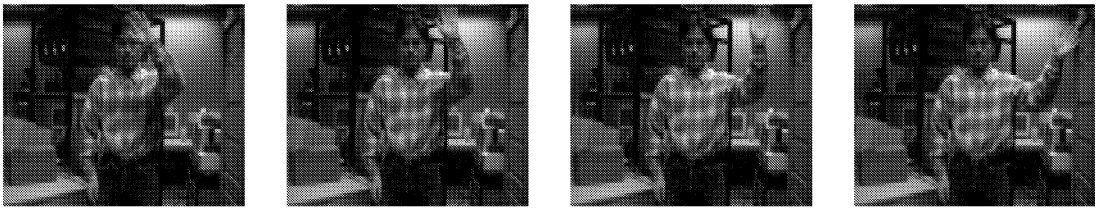
以上に述べてきた問題への具体的な取り組みのために，本研究では新たに身振りプロトコルの概念を導入する．本論文では，身振りプロトコルを『ある身振りを認識しようとする際に，何（着眼点や特徴）をどの程度の広がり（視野や座標系）をもって認識するのかをあらかじめ誘導する経験的な取り決め』と定義して，認識系を身振りプロトコルに適応させるための課題に取り組む．身振りプロトコルは，身振り画像中に含まれる無用な情報の影響を抑えて，より的確に身振りを認識するための極めて重要な役割を担っている．

身振りプロトコルの学習が可能となれば，利用者が独自に定義する身振りの認識のみならず，身振り画像に含まれる無用な情報の影響を抑えた，より頑健な身振り認識処理の実現が期待できる．



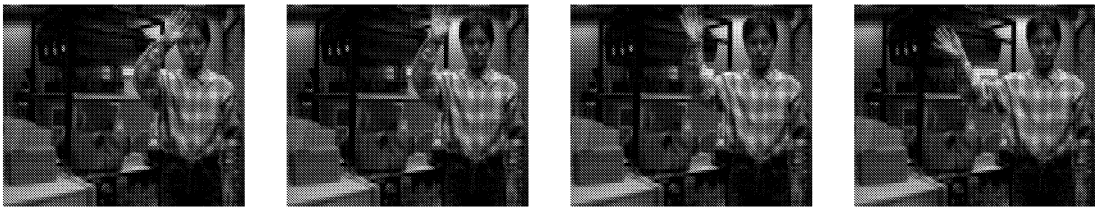
＃１　　＃２　　＃３　　＃４

身振り動作（A）



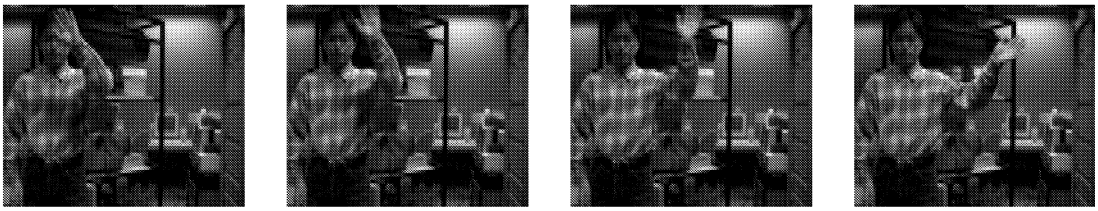
＃１　　＃２　　＃３　　＃４

身振り動作（B）



＃１　　＃２　　＃３　　＃４

身振り動作（C）



＃１　　＃２　　＃３　　＃４

身振り動作（D）

図 1.1　身振り画像の特徴選択問題

1.4.2 身振り画像のサンプリング問題

身振り認識処理では，まず，人物の動作領域を実画像中から抽出する必要がある．「利用者の意向は身体動作に表出される」と想定できるならば，画像の時間差分操作により利用者の意向を反映した動作領域を実画像中から容易に抽出できるはずである．画像差分処理により動作領域を抽出する場合，画像全体をテンプレートマッチング手法により走査・照合する方式に比べて，処理の大幅な簡略化と高速化を図れるメリットがある．

問題は，認識処理の対象となる図形パターンの位置や大きさそして回転といった基本要素について先験的な制限を設ける明確な根拠が存在しないことにある．この点において身振り認識問題は，目，鼻，口などを「部品」という概念で表現し，能動的なオブジェクト探索を行う従来の顔画像認識問題とは異なっている．もし，身振りに関係する身体部位を「部品」として表現する場合，身体のどこからどこまでの部位を「部品」とするかを事前に明らかにする必要があるが，その数や組み合わせは膨大であり，現実的でない．

このように，身振り認識問題では身体の部品表現のアイディアは必ずしも最適ではなく，身振り認識問題に特化した図形パターンの表現方法と画像差分処理に基づく領域分割手法が要求されていることは明らかである．しかしながら，画像差分操作を適用する場合，画像取得の際のサンプリング間隔が一定でないと，結果として得られる図形パターンとその領域情報の信頼性が低下してしまう問題がある．このため，画像差分操作により動作領域を抽出する方式では，画像サンプリング間隔の安定化を図ることが不可避の課題となる．

信頼性の高い画像差分操作の下で動作領域を抽出するには，身振り画像のサンプリング問題に取り組めばよい．画像サンプリング問題に取り組むことにより，任意の速度で身振りを認識するシステムの実現も期待できる．サーボ機構の制御を目的とした視覚フィードバック制御の研究例 [38] はあるが，従来の身振り画像認識手法において制御パラダイムを導入している例はほとんど見当たらず，静止画の画像処理手法をそのまま適用する例がほとんどである．より実用的な仮想現実感のための身振りインタフェースの実現には，身振り画像のサンプリング問題に取り組むことが不可欠である．

1.4.3 身振り情報の実時間推定

従来の身振り画像認識研究では，認識結果をそのままジェスチャコマンドとして利用するに留める例が多いなど，身振りをコマンド入力のための単なる一手段としてしか捉えていない傾向が強い．

身振りによる指示入力の有効性を高めるためには，入力された身振り画像のクラス推定のみでなく，身振り動作の向きや相対的な速度（以降，身振り速度と呼ぶ），相対的な振幅（以降，身振り振幅と呼ぶ）などの複数のパラメータを同時に推定し，仮想現実感システムに実時間で反映させることが重要である．図 1.2 に本研究が目指す身振りインタフェースの概念図を示し，図 1.3 に推定の対象とする身振り情報を示す．身体動作により伝達されるアナログ的な情報は，いわゆる感性情報 [39] に属するが，本論文ではこれを「身振り情報」と呼ぶことにする．

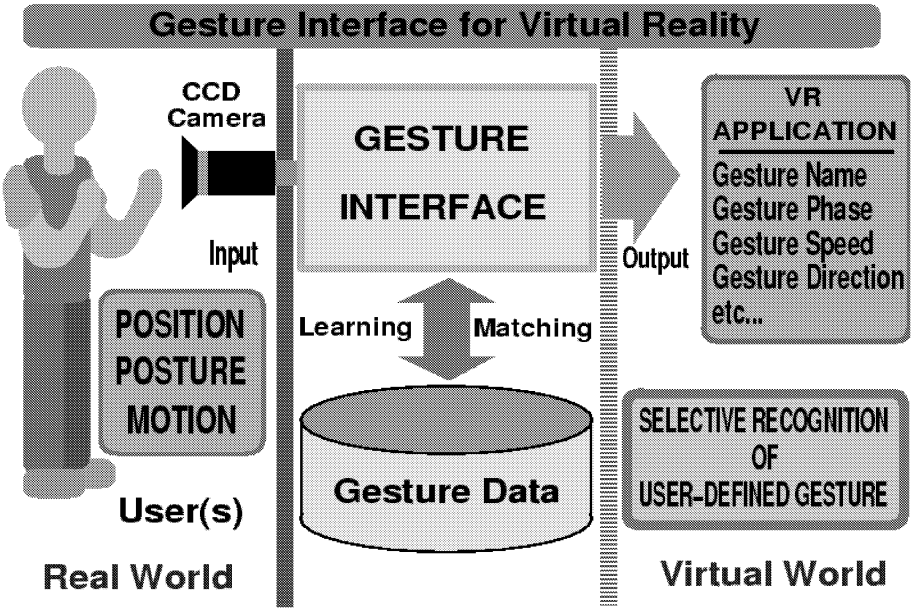


図 1.2 仮想現実感のための身振りインタフェース

図 1.2 に示すように，ＣＣＤカメラにより撮影された利用者の身振り画像は画像入力装置によりコンピュータに入力される．続いて，身振りインタフェースにより身振り情報が認識され，その結果は仮想現実感システムへと引き渡される．

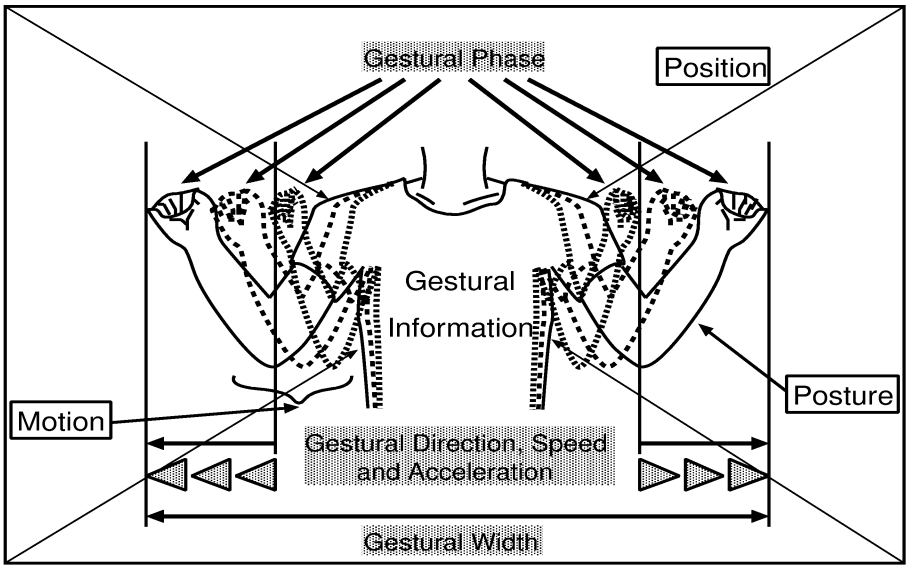


図 1.3 実時間推定の対象とする身振り情報

本論文では，認識の対象とする身振りの種類や性質を限定せず，利用者により独自に定義される身振りの認識をも可能とする身振りインタフェースの実現手法を提案する．

2 多注視点身振り認識法

同一種類の身振りとして定義される動作には，何らかの視覚的な共通項が存在することが多い．我々は，こうした視覚的な共通項を選択的に注視することで，たとえ雑踏の中にあったとしても，コミュニケーションの相手が示す身振りの意味を的確に読み取っている．

こうした能力は，仮想現実感環境における利用者の身振り動作を読み取る際にも不可欠である．しかし，身振りの認識を可能とする視覚的共通項は，常に自明であるとは限らない．時には，複数の視覚的共通項を同時に考慮する必要が生じることもある．こうした能力をコンピュータに付与するには，同一種類として与えられる身振りの画像列から，注視すべき視覚的な共通項を自ら見出す機能の実現が不可欠であり，このことは前述の身振りプロトコル問題に対応するための手法が強く要求されていることを意味する．

本章では，身振りプロトコルの学習とそれに基づく認識を実現するために，多注視点身振り認識法（QVIPS–Quadruple Visual Interest Point Strategy）を提案する．図 2.1 に提案手法の流れ図を示す．提案手法は，特徴量に基づく照合処理，活性化マップによる特徴統合処理，身振りプロトコルに基づく認識処理の 3 段階により構成される階層型身振り認識機構である．

時々刻々と入力される身振り画像は，複数種類の特徴抽出フィルタにかけられ，その結果，複数注視点に対応する特徴量が抽出される．抽出された特徴量は，学習時に身振り標準パターンとして登録される．一方，認識時には身振り標準パターンとの照合処理が行われ，照合結果は活性化マップとして出力される．この段階での処理を「特徴量に基づく照合処理」と呼ぶ．

各注視点の重み付けのために，活性化マップを利用した身振りプロトコルの学習（以降，プロトコル学習と呼ぶ）が行われる．プロトコル学習では，ある身振

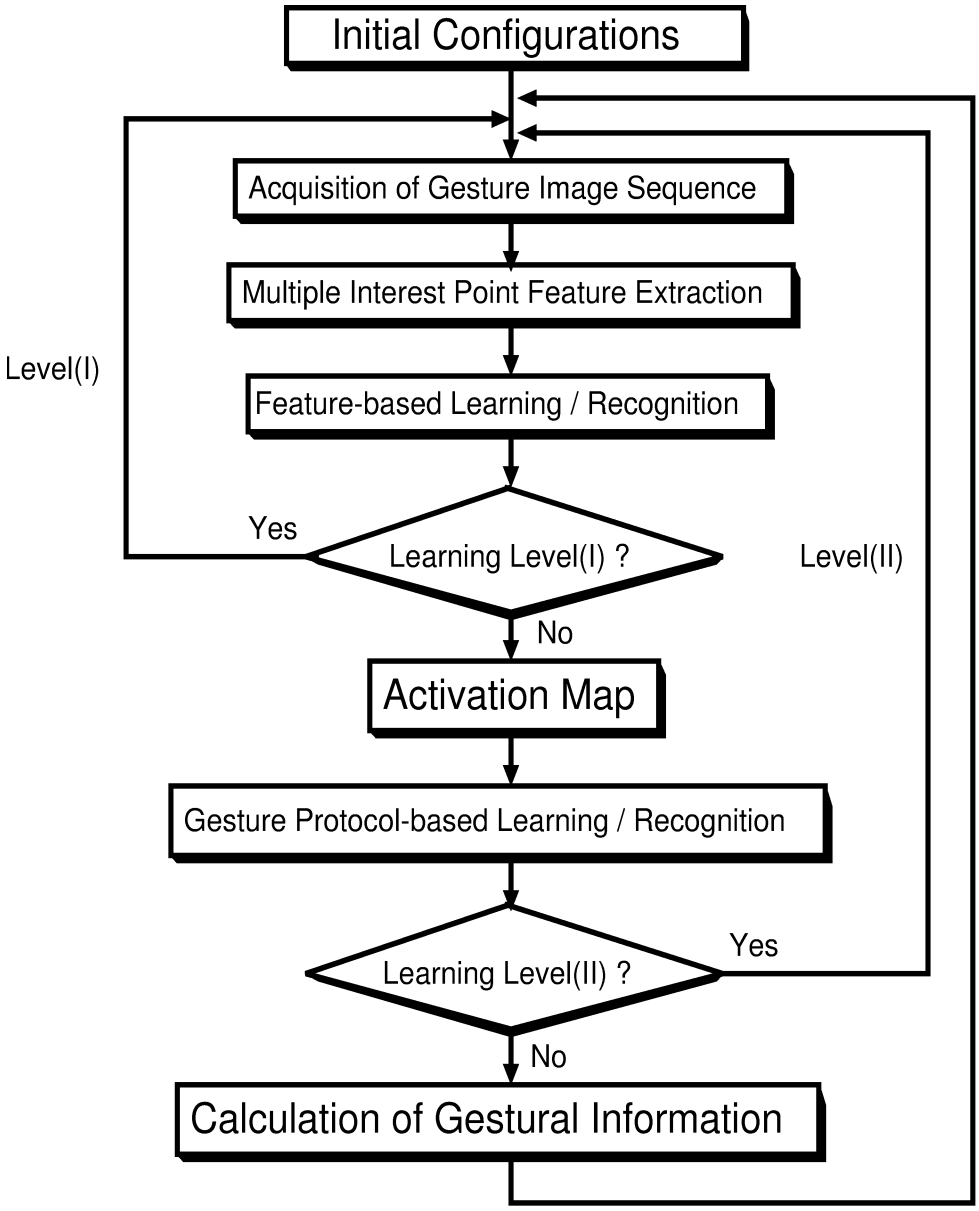


図 2.1 多注視点身振り認識法の流れ図

りを認識する際に時間領域で安定している注視点に，相対的に大きな重みが割り当てられる．さらに，各注視点に対応するゆう度分布（以降，プロトコルマップと呼ぶ）を生成・登録し，以後，このプロトコルマップに基づいた認識処理が行われる．この段階での処理を「身振りプロトコルに基づく認識処理」と呼ぶ．

入力された身振り画像のクラス推定の後，活性化マップに基づいて身振り情報が推定される．推定される身振り情報は，標準画像列のフレーム番号に対応する身振り位相値，さらに標準画像列と比べた場合の身振り速度や身振り振幅である．図 2.2 に提案手法の基本的枠組みを示す．

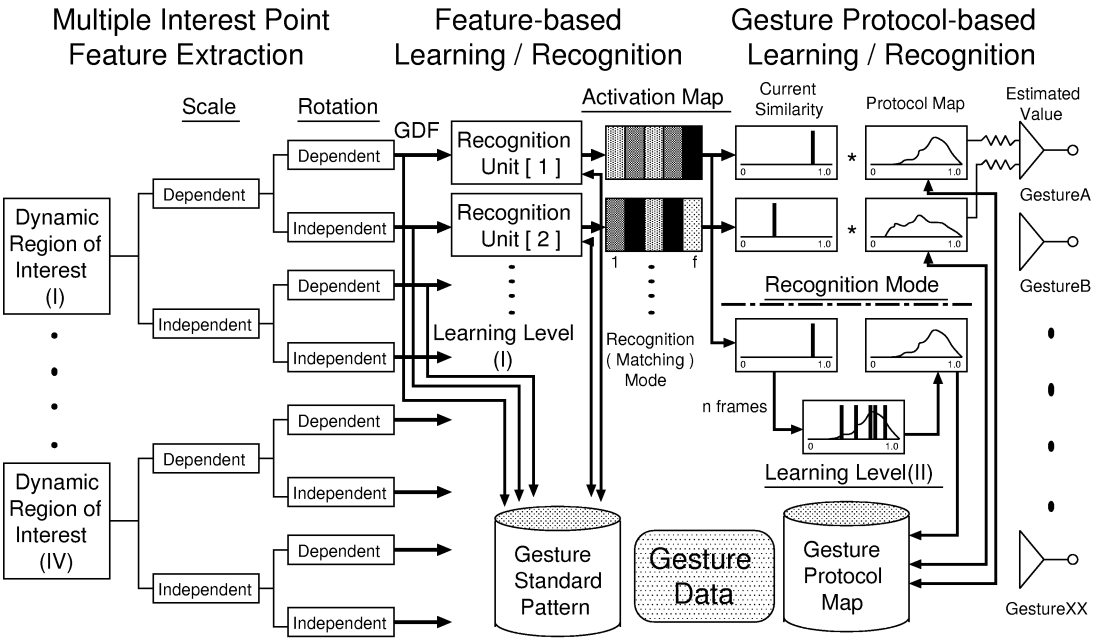


図 2.2　多注視点身振り認識法の基本的枠組み

2.1 身振り画像の動的注視

身振りは視覚的な特徴群から構成されていると考えられるが，身体動作から視覚的特徴を抽出するためには，まず，注視領域が設定されている必要がある．我々は普段，動作変化の生じている領域に注意を傾ける一方で，変化しない領域に対しても無意識に注意を払っている．もし，我々の視覚系が，動作変化だけに注意を向けて身振りを観測するのであれば，相手の姿を意識することなく，単に動作部位のパターン変化のみを我々は目前に見ることになるであろう．実際には，相手の姿を意識しながら，動作部位にも注意を払っていることは明らかであり，我々は，身振りを観測するとき，動作変化のある領域と変化の乏しい領域の双方に同時的注意を傾けている．

ネコやサルの網膜神経節細胞には，空間情報の伝送を分担するX形と，時間情報の伝送を分担するY形の二種類が存在するとの指摘がある [40]．究極的な身振り認識システムは，人間の視覚系そのものであり，このことは，人間の視覚系に倣った処理方式を実現する必要があることを示唆している．そこで，差分画像をY形細胞，シルエット画像をX形細胞に対応させることにより，生体の視覚系に倣った身振り認識システムの構築を試みる．具体的に提案手法では，動作の変化を検出するために差分画像を生成し，動作の変化に乏しい領域を検出するためにシルエット画像を生成する．

単純図形における注視領域，および，注視点の設定に関する理論的解析は，飯島により行われている [41] が，身振り画像のような複雑な図形パターンにおける注視領域，および，注視点の設定方法については，ほとんど研究されていない．このように身振り画像における注視領域の設定は従来から困難であり，本手法では，差分画像とシルエット画像の領域情報と形状パターンを組み合わせることによりフレーム毎の注視領域を設定する．

図 2.3 に身振り画像の動的注視処理の流れを示す．差分画像は，動作部位の姿勢とその重心位置（動的注視領域 II），さらに人物中心座標系から見た動作部位の相対的な位置とその形状（動的注視領域 I）に関する特徴量を抽出するために利用する．シルエット画像は，人物の位置（動的注視領域 IV）や姿勢（動的注視領域 III）に関する特徴量（重心位置，範囲，形状）を抽出するために利用する．図 2.3 に示すように，シルエット画像と差分画像および，それらから得られる重

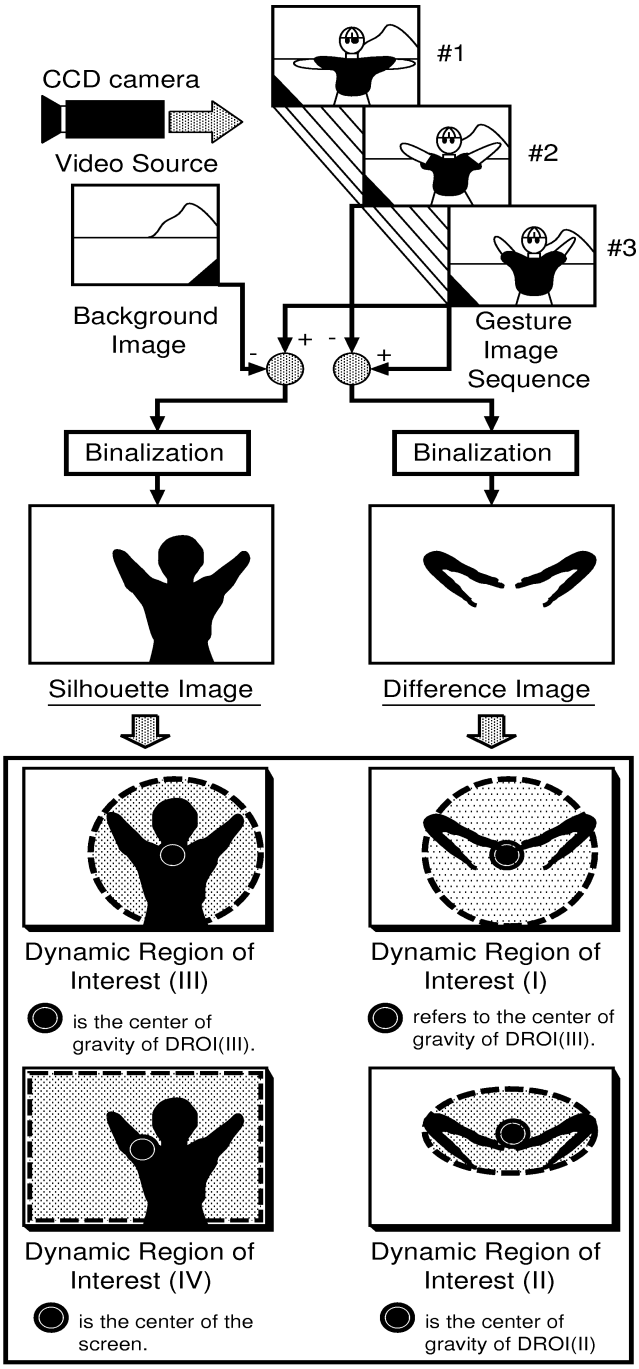


図 2.3 身振り画像の動的注視手法

心位置（動的注視点）や領域情報を組み合わせることにより身振り画像を動的に注視し，利用者の位置・姿勢・動作に関する特徴量をフレーム毎に抽出する．

2.2 動的注視領域からの形状特徴抽出

提案手法は多注視点方式であるため実時間での身振り認識処理を実現するには，形状特徴の抽出・照合処理を効率よく高速に行わなければならない．そこで，身振りに適した形状特徴の抽出手法が必要になるが，本節では，二次元図形パターンとガウス分布との畳み込み演算を行い，図形パターンを一次元の形状特徴パターンに変換する手法を提案する．

動的注視領域における身振り画像を $f_\tau(x,y)(\tau = 1,2,...,N)$ とする．閾値 t により身振り画像 $f_\tau(x,y)$ を２値化し，得られる２値画像を $g_\tau(x,y)$ とする．$g_\tau(x,y)$ に対する原点回りのモーメント $m_{p,q}$ は式 (2.1) で定義される．

$$m_{p,q} = \sum_x \sum_y g_\tau(x,y) x^p y^q \tag{2.1}$$

注視点 (a_τ, b_τ) を式 (2.2) により定義する．

$$(a_\tau, b_\tau) = \left(\frac{m_{1,0}}{m_{0,0}}, \frac{m_{0,1}}{m_{0,0}} \right) \tag{2.2}$$

ここで $g_\tau(x,y)$ を，(a_τ, b_τ) が原点である極座標系 $P_\tau(r,\theta)$ で表現する．角度 θ における半径 $(= R)$ 方向のパターン分布を Ω 分割する．個々の分割領域をカーネルと呼び，その角度 θ における形状特徴量 $s_\varepsilon(\theta)(\varepsilon = 1,2,...,\Omega)$ を式 (2.3) により定義する．

$$s_\varepsilon(\theta) = \frac{R}{\Omega} \frac{\sum_r P_\tau(r,\theta) \exp\left\{-a(r-\phi)^2\right\}}{\sum_r P_\tau(r,\theta)} \tag{2.3}$$

ただし，a は形状特徴量のユニークさに関係する勾配係数であり，ϕ は位相項である．a と ϕ については付録A（→形状特徴抽出法）で詳しく述べる．式 (2.3) を任意の解像度で２値画像の全周方向に適用することによりカーネル ε における形状特徴パターン $s_\varepsilon(\theta)$ を得る．式 (2.3) を適用して得られる形状特徴パターンは，各カーネルの外郭部形状をより反映したものとなる．対象パターンとガウス分布

表 2.1 設定した１６通りの注視点

注視点番号	位置	大きさ	回転	特徴情報源
1	依存	依存	依存	動的注視領域 I
2	依存	依存	非依存	動的注視領域 I
3	依存	非依存	依存	動的注視領域 I
4	依存	非依存	非依存	動的注視領域 I
5	非依存	依存	依存	動的注視領域 II
6	非依存	依存	非依存	動的注視領域 II
7	非依存	非依存	依存	動的注視領域 II
8	非依存	非依存	非依存	動的注視領域 II
9	非依存	依存	依存	動的注視領域 III
10	非依存	依存	非依存	動的注視領域 III
11	非依存	非依存	依存	動的注視領域 III
12	非依存	非依存	非依存	動的注視領域 III
13	依存	依存	依存	動的注視領域 IV
14	依存	依存	非依存	動的注視領域 IV
15	依存	非依存	依存	動的注視領域 IV
16	依存	非依存	非依存	動的注視領域 IV

との畳み込みを計算し，その結果に密度係数を掛け合わせることから，$s_\varepsilon(\theta)$ をガウス密度特徴（GDF–Gaussian Density Feature）と以降呼ぶ．

ガウス密度特徴を求めた後，各カーネルにおける形状特徴パターン $s_\varepsilon(\theta)$ に高速フーリエ変換（FFT）を適用する．パワースペクトルのシフト不変の性質から，FFT の結果得られるパワースペクトル $P_\varepsilon(\omega)$ を回転不変の形状特徴パターンとして利用する．$s_\varepsilon(\theta)$ 自体は回転依存の特徴パターンとして利用する．

提案手法では以下に示す４通りの処理を各動的注視領域に適用することにより表 2.1 に示す１６通りの注視点に対応する形状特徴パターンを抽出する．

［１］$s_\varepsilon(\theta)$ を正規化することにより，回転に依存するが大きさには依存しない

形状特徴パターンを得る．

［２］$s_\varepsilon(\theta)$ を正規化しない場合，回転と大きさに依存する形状特徴パターンを得る．

［３］$P_\varepsilon(\omega)$ を正規化することにより，回転と大きさに依存しない形状特徴パターンを得る．

［４］$P_\varepsilon(\omega)$ を正規化しない場合，回転には依存しないが大きさには依存する形状特徴パターンを得る．

2.3 特徴量に基づく学習処理

提案手法では，各身振り画像から抽出される１６通りの注視点に対応した形状特徴パターンと身振り画像のフレーム番号（身振り位相値）を身振り標準パターンとして登録する．パターン番号 g，注視点番号 l の身振り標準パターンに登録されている形状特徴パターンを

$$\boldsymbol{K}_l^{(g)} = (\boldsymbol{K}_1, \boldsymbol{K}_2, \cdots, \boldsymbol{K}_\xi, \cdots, \boldsymbol{K}_\Theta)^t$$

と表す．ここで，

$$\boldsymbol{K}_\xi = (\boldsymbol{s}_1(\boldsymbol{\theta}), \boldsymbol{s}_2(\boldsymbol{\theta}), \cdots, \boldsymbol{s}_\Omega(\boldsymbol{\theta}))^t \quad , \quad \boldsymbol{\theta} = \tfrac{2\pi\xi}{\Theta}$$

である．本手法により一枚の身振り画像を分解表現することで，プロトコル学習における注視点の重み付け処理が可能となる．身振り標準パターンの一貫性を保証するために，各身振りクラスにつき一動作分の標準パターンを登録する．

2.4 特徴量に基づく照合処理

本節では図 2.2 中の認識ユニットにおける照合処理について述べる．現在入力されている身振り画像の注視点 l における形状特徴パターンを

$$\boldsymbol{T}_l = (\boldsymbol{T}_1, \boldsymbol{T}_2, \cdots, \boldsymbol{T}_\xi, \cdots, \boldsymbol{T}_\Theta)^t$$

とし，各注視点に対応する認識ユニットに入力した際の $\boldsymbol{T}_l$ と身振り標準パターン $\boldsymbol{K}_l^{(g)}$ との距離 $d_l^{(g)}$ を式 (2.4) により定義する．

$$d_l^{(g)} = \sum_{\xi=1}^{\Theta} \| \boldsymbol{T}_\xi - \boldsymbol{K}_\xi \| \tag{2.4}$$

ここで Θ は角度分解能に相当する解像度，g は任意の身振り標準パターン番号である．身振りクラス i の身振り標準パターンにおいて $d_l^{(g)}$ を最小とする身振り標準パターンの番号を k_i とすると，注視点 l，身振りクラス i におけるフレーム番号 f の入力身振り画像に対する類似度 $S_{lf}^{(i)}$ は式 (2.5) により定義することができる．

$$S_{lf}^{(i)} = 1 - \frac{d_l^{(k_i)}}{\mathrm{Max}\left(d_l^{(g)}\right)} \tag{2.5}$$

ここで Max() は最大値を返す関数である．式 (2.5) により得られる類似度 $S_{lf}^{(i)} (0 \leq S_{lf}^{(i)} \leq 1)$ とパターン番号 k_i の身振り標準パターンに格納されている位相値を特徴量に基づく照合処理の結果として，次節で述べる身振りプロトコルに基づく学習・認識処理部に引き渡す．

引き渡される照合結果は，各注視点における類似度の多次元時系列パターン $\boldsymbol{X}_l^{(i)} = (\boldsymbol{S}_{l_1}^{(i)}, \boldsymbol{S}_{l_2}^{(i)}, \cdots, \boldsymbol{S}_{lK}^{(i)})^t$ となる．特徴量に基づく照合処理の結果を各注視点における類似度 $S_{lf}^{(i)}$ に変換する理由としては，各注視点で得られるパターン間距離が大きさや回転などの異なる評価尺度であり，照合結果を統一することによりすべての注視点を等価にすることが挙げられる．

2.5 身振りプロトコルに基づく学習・認識処理

特徴量に基づく照合処理では空間的類似度のみを考慮するため，時間軸方向の類似度の変動は考慮していない．ある身振りを認識する際に，時間軸方向に安定した類似度を与える注視点は重要な手掛かりである．従って，空間的類似度の時間的変動を考慮することは身振りの学習において極めて重要である．

身振りクラス i，注視点 l における活性化マップが，$\boldsymbol{X}_l^{(i)} = (\boldsymbol{S}_{l_1}^{(i)}, \boldsymbol{S}_{l_2}^{(i)}, \cdots, \boldsymbol{S}_{lK}^{(i)})^t$ により与えられている場合，$\boldsymbol{X}_l^{(i)}$ の平均 $\mu_l^{(i)}$ ，分散 $\sigma_l^{(i)}$ は，

$$\mu_l^{(i)} = \frac{1}{K} \sum_{k=1}^{K} \boldsymbol{S}_{lk}^{(i)} \tag{2.6}$$

$$\sigma_l^{(i)} = \frac{1}{K}\sum_{k=1}^{K}\left(\boldsymbol{S}_{lk}^{(i)} - \mu_l^{(i)}\right)\left(\boldsymbol{S}_{lk}^{(i)} - \mu_l^{(i)}\right)^t \tag{2.7}$$

により与えられる．ここで注視点重み $\omega_l^{(i)}$ を式 (2.8) により定義する．

$$\omega_l^{(i)} = \frac{\mu_l^{(i)}}{\sigma_l^{(i)} + \alpha} \tag{2.8}$$

α を注視点強調係数とし，α が小さい程分散 $\sigma_l^{(i)}$ が $\omega_l^{(i)}$ に大きな影響を与えるようにする．また，類似度が安定している程 $\sigma_l^{(i)}$ は小さくなり，$\omega_l^{(i)}$ は大きく設定される．プロトコル学習の際には，活性化マップ $\boldsymbol{X}_l^{(i)}$ をプロトコルデータ

$$\boldsymbol{M}_l^{(i)} = (\boldsymbol{M}_{l_1}^{(i)}, \boldsymbol{M}_{l_2}^{(i)}, \cdots, \boldsymbol{M}_{lM}^{(i)})^t$$

として登録するのみでなく，各注視点における注視点重み $\omega_l^{(i)}$ を算出し登録しておく．

一方，身振りプロトコルに基づく認識処理では，プロトコルデータ $\boldsymbol{M}_l^{(i)}$ の各要素の分布型がガウス分布に従うものと仮定し，フレーム番号 f，活性化マップ $\boldsymbol{X}_l^{(s)}$ の身振り画像が入力された場合の評価値 $E_f^{(i)}$ を，

$$E_f^{(i)} = \sum_{l=1}^{L}\sum_{m=1}^{M}\exp\left\{-\beta\left(S_{lf}^{(s)} - M_{lm}^{(i)}\right)^2\right\} \tag{2.9}$$

により定義する．β は分離係数であり，β が大きい程類似度要素間の影響が小さくなる．

ここで，注視点 l において式 (2.9) に従いすべての類似度区間 $\lambda[0,1]$ におけるゆう度を求めた分布を，注視点 l のプロトコルマップとする．プロトコルマップは各身振りクラスにおいて生成する．プロトコルマップはプロトコル学習の直後，すなわち，$\boldsymbol{M}_l^{(i)}$ が確定した後に式 (2.10) に基づき一括して生成できるため，実装の際には演算結果のテーブル化により評価値計算の高速化を図る．

$$I_{l,\lambda}^{(i)} = \sum_{m=1}^{M}\exp\left\{-\beta\left(\lambda - M_{lm}^{(i)}\right)^2\right\} \tag{2.10}$$

なお，本手法では各身振りクラスのプロトコルマップを独立して生成するため，原理的に認識対象とする身振りクラスの数に制限はない．

さて，N フレームの身振り画像系列が入力された際の累積評価値 $E^{(i)}$ を，

$$E^{(i)} = W_i \sum_{f=1}^{N} E_f^{(i)} \tag{2.11}$$

により定義する．ここで W_i は各身振りクラスに与える重みであり，全注視点数を L とした場合，

$$W_i = \frac{L}{\sum_{l=1}^{L} \omega_l^{(i)}} \tag{2.12}$$

により定義する．式 (2.12) より，プロトコル学習の結果，注目すべき注視点が定まっている程，クラス重み W_i は大きく設定される．なお，動作開始から現在までの評価値 $E_f^{(i)}$ を加算した累積評価値 $E^{(i)}$ が最大になる身振りクラス C に，入力された身振り画像は属するものと判定する．

2.6 身振り情報の定量化

入力された身振り画像のクラス推定の後，活性化マップ $X_l^{(C)}$ に格納されている身振り位相値の時系列変化から，身振り標準パターンと比較した場合の身振り位相値・方向・身振り速度・身振り振幅を定量化する．

2.6.1 身振り位相値

身振りクラス C の注視点 l における活性化マップに格納されている位相値の時系列パターンを $\boldsymbol{P}_l^{(C)} = (\boldsymbol{P}_{l_1}^{(C)}, \boldsymbol{P}_{l_2}^{(C)}, \cdots, \boldsymbol{P}_{lK}^{(C)})^t$ とする．今，フレーム番号 k における入力身振り画像の推定位相値 P_{0k} を，

$$P_{0k} = \frac{\sum_{l=1}^{L} P_{lk}^{(C)} Z_{lk}}{\sum_{l=1}^{L} Z_{lk}} \tag{2.13}$$

により推定する．ここで Z_{lk} はフレーム番号 k，注視点 l における評価値重みであり，

$$Z_{lk} = \sum_{k=1}^{K} \sum_{m=1}^{M} \exp\left\{-\beta \left(S_{lk}^{(C)} - M_{lm}^{(C)}\right)^2\right\} \tag{2.14}$$

により与える．β は前節で述べた分離係数である．位相値 P_{0k} の急激な変動を抑制するために，式 (2.15) に従いフィルタ長 U の移動平均処理を適用した結果得

られる P_k を入力された身振り画像の推定身振り位相値とする．

$$P_k = \frac{\sum_{u=0}^{U-1} P_{0k-u}}{U} \tag{2.15}$$

2.6.2 身振り速度と身振り振幅

身振り速度 s と身振り振幅 w を式 (2.16) と式 (2.17) により算出する．

$$s = P_k - P_{k-\triangle} \tag{2.16}$$

$$w = \frac{\mathrm{Max}(P_k) - \mathrm{Min}(P_k)}{Fmax^{(C)}} \times 100(\%) \tag{2.17}$$

ここで Max() は最大値を返す関数，Min() は最小値を返す関数である．また，式 (2.16) の $\triangle$ は，微小時間幅に対応するフレーム数，式 (2.17) の $Fmax^{(C)}$ は身振りクラス C の標準画像列のフレーム数である．式 (2.16) より分かるように登録時の身振り動作の方向を順方向とすれば，順方向の身振り動作の際 s は正の値，停止の際には 0，逆方向身振り動作の際には負の値をとる．

2.7 評価実験

提案手法をワークステーションに実装し評価実験を行った．ＣＣＤカメラにより撮影された画像は画像入出力装置 Galileo Video を通してワークステーション（SGI Indigo2）に解像度横 160[dot] 縦 120[dot] で取り込まれ，オンライン認識される．実験システムの構成を図 2.4 に示す．

なお，すべての処理をソフトウェアで行っている．また，特別な照明や背景は使用せず通常の実験室内で行った．参考のために，本実験の対象とした身振り動作のスナップショット画像を付録Ｂに示す．本実験では，勾配係数 $a = 5.0$，カーネル数 $\Omega = 1$，注視点強調係数 $\alpha = 0.1$，分離係数 $\beta = 2000$，移動平均フィルタ長 $U = 3$，微少時間幅フレーム数 $\triangle = 5$ と設定して評価した．

図 2.5 に評価実験システムの実行画面例を示す．身振りの登録および学習は，評価実験用ＧＵＩを利用する．評価実験用ＧＵＩは，システムの内部状態を表示するウィンドウ群，ビデオ画像を表示するウィンドウ，および，認識結果を表示するウィンドウ群から構成される．身振り学習の際には，数秒の準備時間の後，被

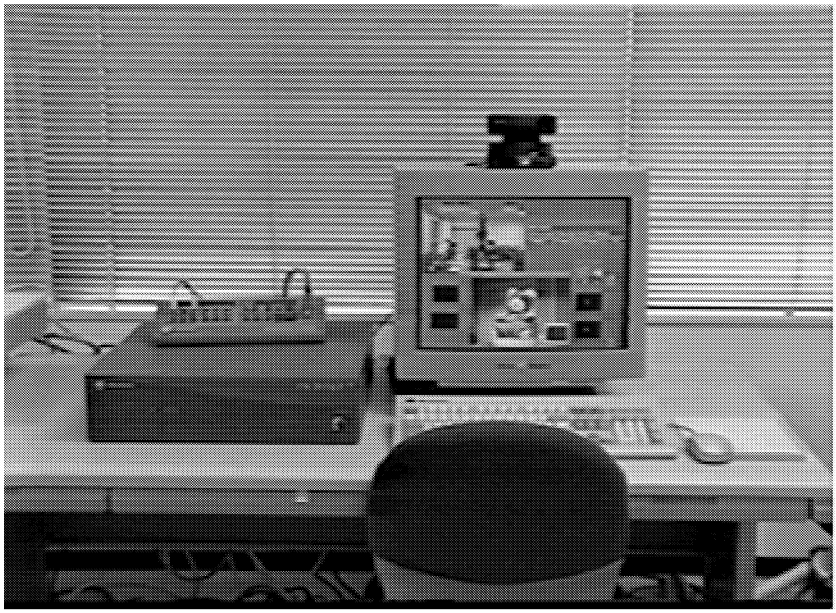


図 2.4 実験システム

図 2.5 実験システムの実行画面例

験者はカメラの前で身振り動作を行う．なお，本論文では，評価実験の信頼性を高めるために，デジタルビデオカメラを使用して身振り画像を入力する．トラッキング情報は，認識結果表示用ウィンドウにフレーム毎に表示される．

評価実験システムでは，以上に述べた機能のみでなく，身振り情報のグラフ表示機能，特徴画像の表示機能，学習データの保存および読出し機能，各種閾値の設定機能を利用することができる．また，プロトコルマップ表示機能，類似度マップ・位相値マップに関する活性化マップの表示機能，ガウス密度特徴の表示機能なども利用でき，実時間でシステムの内部状態を視覚的に確認できる．

2.7.1 身振りプロトコルの学習実験

本節では身振り「バイバイ」と「グッパ」について，身振りプロトコルの学習実験を行った際の結果を報告する.「バイバイ」の身振り動作は「右手を振っても左手を振っても良い」とする一方,「グッパ」の身振り動作は「どこで手のひらを開いても良い」として，類似身振りサンプルを作成し，これらから身振りプロトコルを提案手法により学習できるかどうかを調べる．

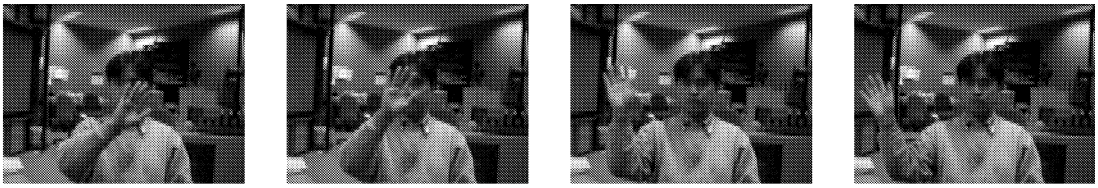
実験は次の３段階に分けて行う．

［１］標準身振りサンプルを用いて身振り標準パターンを学習させる（右手の動作について１回）．

［２］類似身振りサンプルを用いて身振りプロトコルを学習させる（右手と左手をそれぞれの動作について交互に１０回）．

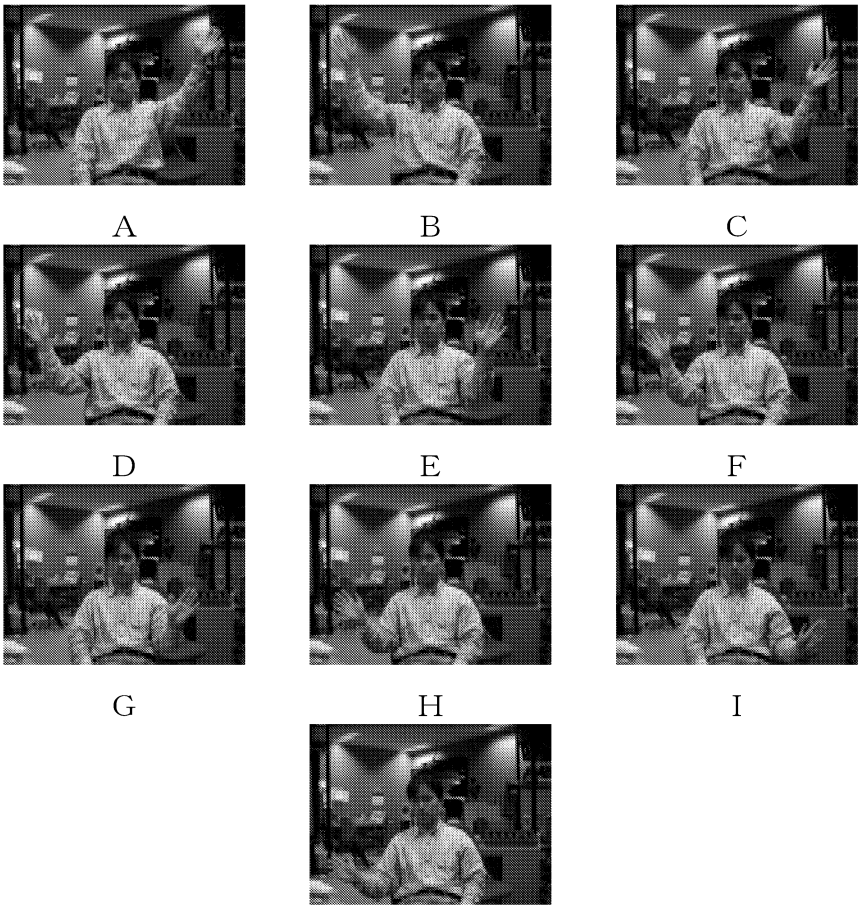
［３］図 2.7 と図 2.9 に示す類似身振りのテストサンプルを使って評価実験を行う．

なお，特徴量に基づく学習で使用した身振り画像のスナップショット例を「バイバイ」に関しては図 2.6 に,「グッパ」に関しては図 2.8 に示す．



＃１　　＃２　　＃３　　＃４

図 2.6　「バイバイ」の標準身振りサンプル



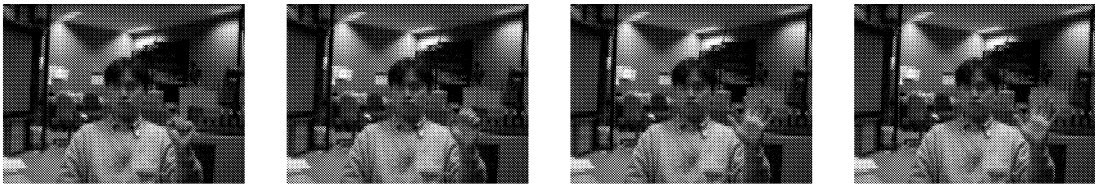
A　　B　　C

D　　E　　F

G　　H　　I

J

図 2.7　「バイバイ」の類似身振りサンプル



＃１　　＃２　　＃３　　＃４

図 2.8　「グッパ」の標準身振りサンプル

A　　B　　C

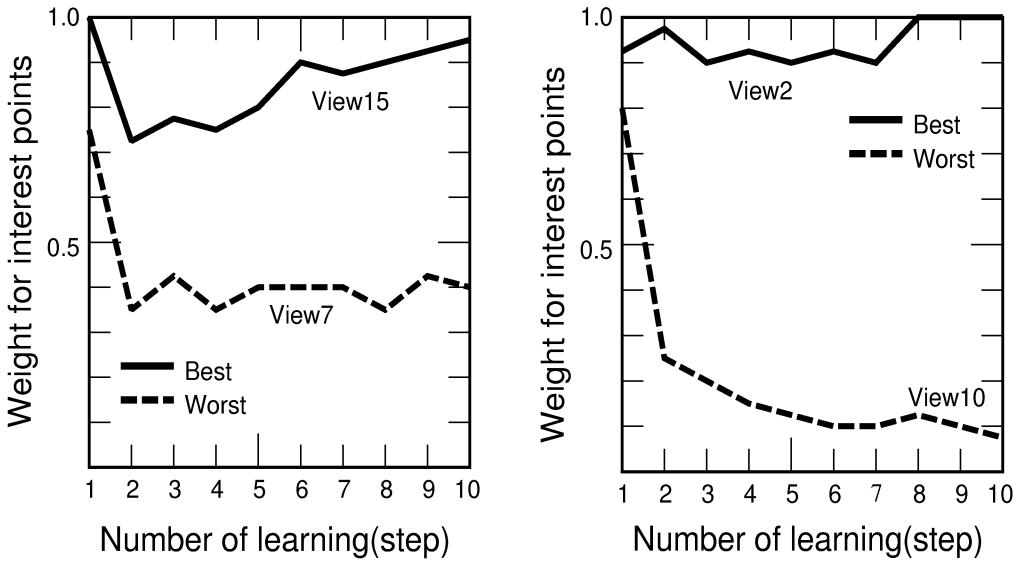
D　　E　　F

G　　H　　I

J

図 2.9　「グッパ」の類似身振りサンプル

身振りプロトコルを学習させた際に得られた注視点重みの変化を図 2.10 に示す．なお，図 2.10 においては全ての注視点重みではなく，最終的に重みが最大になる注視点と最小になる注視点に関する重みの変化を示した．



「バイバイ」　　「グッパ」

図 2.10　注視点重みの変化

プロトコル学習の後，次の６種類の注視点条件の下で身振り認識実験を行った．

［条件１］差分画像で最も有効な注視点のみ

［条件２］シルエット画像で最も有効な注視点のみ

［条件３］すべての注視点の中で最も有効な４つの注視点

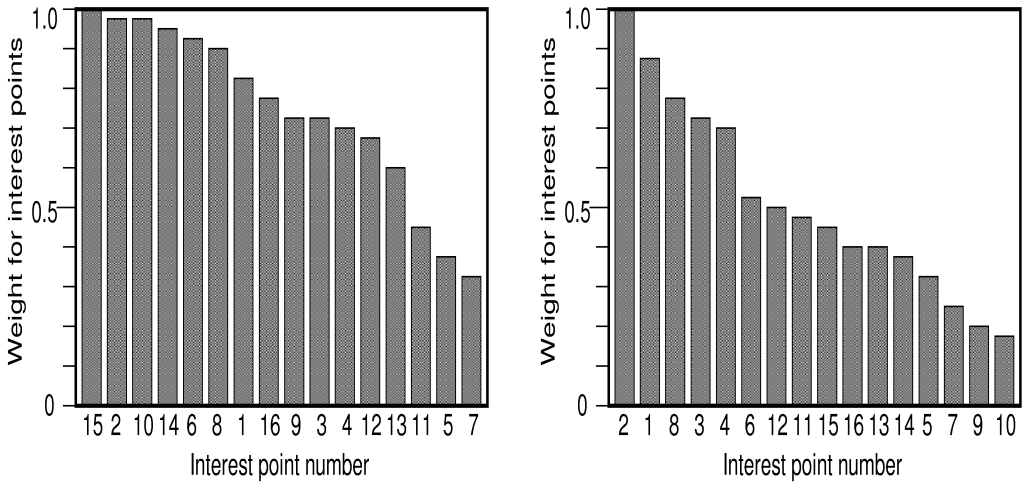
［条件４］差分画像に関するすべての注視点

［条件５］シルエット画像に関するすべての注視点

［条件６］すべての注視点

ここで条件３の「すべての注視点の中で最も有効な４つの注視点」とは，プロトコル学習の結果として得られる注視点重みの中で重みの大きい順に４つの注視点を選択することを意味する．従って，どの注視点が選択されるかは事前には予測できず認識対象に依存する．今回の実験では,「バイバイ」の場合，注視点｛１５，２，１０，１４｝が選択され,「グッパ」の場合は注視点｛２，１，８，３｝が選択された．

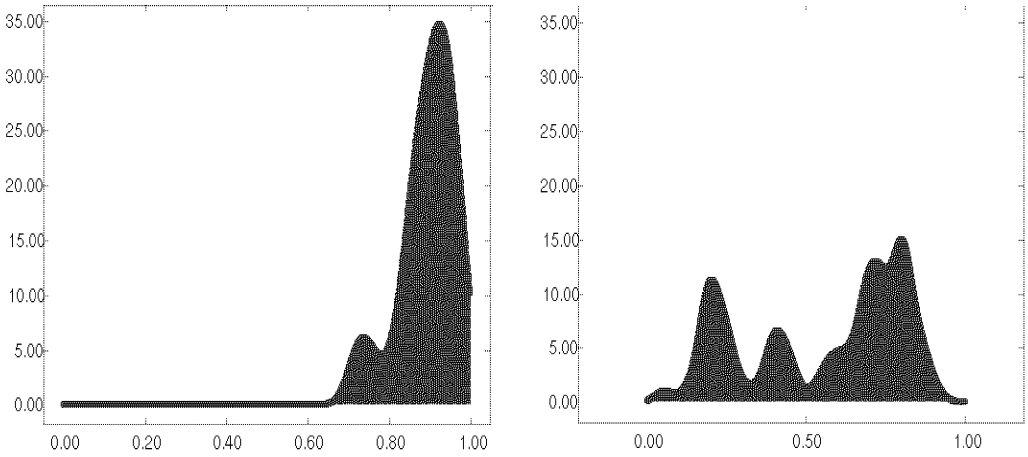
プロトコル学習の結果,「バイバイ」と「グッパ」に関して得られた注視点重みを図 2.11 に示す．また,「バイバイ」において注視点重みが最大となる注視点１５のプロトコルマップと，注視点重みが最小となる注視点７のプロトコルマップを図 2.12 に示す．



「バイバイ」　　「グッパ」

図 2.11　注視点重みの学習結果

６種類の注視点利用条件の下で「バイバイ」の認識実験を行った際の結果を図 2.13 に示す．条件３と条件６の実験条件ですべてのテストサンプルが正しく認識されている．それら以外では差分画像に関する注視点での認識結果である条件１と条件４に良好な認識結果が得られている．これは「身振り“バイバイ”を認識する際には差分画像特徴に注目すれば良い」という常識的な予測を支持する結果である．



注視点１５　　注視点７

図 2.12　プロトコルマップの様子

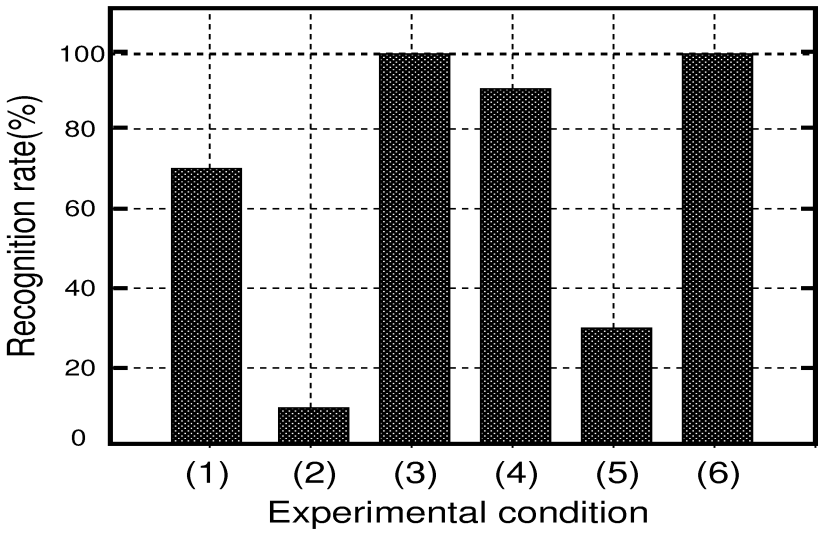


図 2.13　「バイバイ」の認識結果

差分画像とシルエット画像の両方を考慮した場合の結果（条件３と条件６）が差分画像のみを考慮した場合の結果（条件１と条件４）を上回っている．この結果は，複数の特徴画像を同時に考慮することにより，単一の特徴画像のみを考慮する手法に比べてより高い認識率が得られることを示している．

同様にして，６種類の注視点利用条件の下で「グッパ」の認識実験を行った際の結果を図 2.14 に示す．図 2.13 と図 2.14 からも分かるように，条件６（→全ての注視点を利用する場合）と条件３（→全ての注視点で最も有効な４つの注視点を利用する場合）において全てのテストサンプルが正しく認識されている．

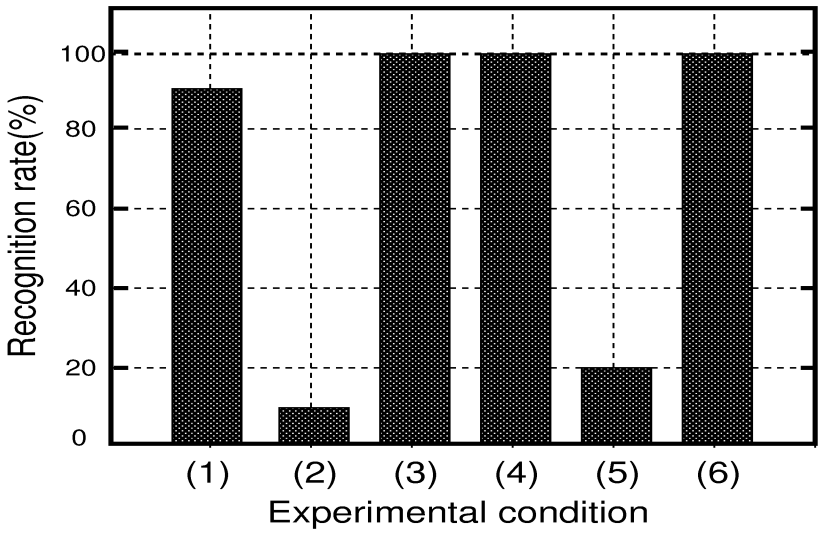


図 2.14 「グッパ」の認識結果

全体的に，時間差分画像に関する注視点による認識結果がシルエット画像の場合よりも良くなっている．これは，プロトコル学習の結果として,「バイバイ」と「グッパ」については「時間差分画像特徴に注目すれば良い」というプロトコルが提案手法により導かれたことを示している．以上の実験結果は，単一の特徴画像のみでは認識が困難な身振りでも，多注視点アプローチに基づくプロトコル学習により，高い認識率が得られることを示している．

2.7.2 衣服依存性と身振り依存性評価実験

提案手法の衣服依存性を評価するために，同一人に図 2.15 に示す 6 種類の衣服Ａ～Ｆを着用させて表 2.2 に示す 7 種類の身振りの認識実験を行った．なお，各身振りのスナップショット画像を付録Ｂ（→評価実験用身振り画像）に示す.

表 2.2 評価実験で用いる身振り

身振り属性 [5]	身振り動作名	省略表現
標識（コード）	「バイバイ」	(G-A)
標識（コード）	「グッパ」	(G-B)
例示子（模倣）	「鳥のまね」	(G-C)
情感表示	「バンザイ」	(G-D)
調整子	「聞き返し」	(G-E)
環境適応子	「腕組み」	(G-F)
オブジェクト適応子	「マウス操作」	(G-G)

衣服 A の標準身振りサンプルを用いて身振り標準パターンを登録した後，衣服 A の（条件A）標準身振りサンプルの場合と（条件Ｂ）類似身振りサンプルの場合の２通りの条件の下で身振りプロトコルを学習させた．この条件の下ですべての類似身振りサンプル（各衣服の各身振りにつき１０通り，合計４２０サンプル）の認識実験を行った際に得られた平均認識率を表すグラフを，衣服（A,B,C,D,E,F）における平均認識率については図 2.16 ，身振り（G-A,B,C,D,E,F,G）における平均認識率は図 2.17 に示す.

図 2.16 において，プロトコル学習の際に（条件Ｂ）の類似身振りサンプルの学習により，各衣服での平均認識率が全体で平均 30(%)（54(%) から 84(%) に）改善されている．一方，図 2.17 において（条件Ａ）での平均認識率の分散が 42.48 であるのに対して（条件Ｂ）では 15.63 であり，身振り依存性が大幅に改善されている．特に，身振り（G-A），(G-B)，(G-G）の改善が顕著である．この結果は，提案手法により身振りの種類や軌跡の大小に依存しない認識処理が実現可能であることを示している．提案システムでは認識対象となる身振りを限定しないため，こうした性質は極めて有効であると考えられる．

衣服Ａ（モザイク模様のシャツ）

衣服Ｂ（白のトレーナー）

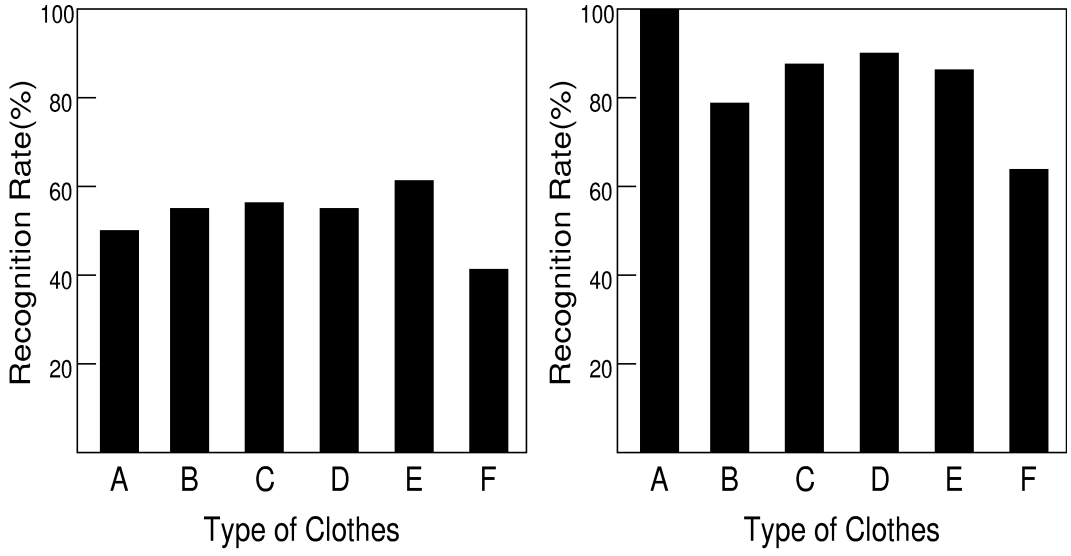
衣服Ｃ（グレーのトレーナー）

衣服Ｄ（縦縞模様のシャツ）

衣服Ｅ（横縞模様のシャツ）

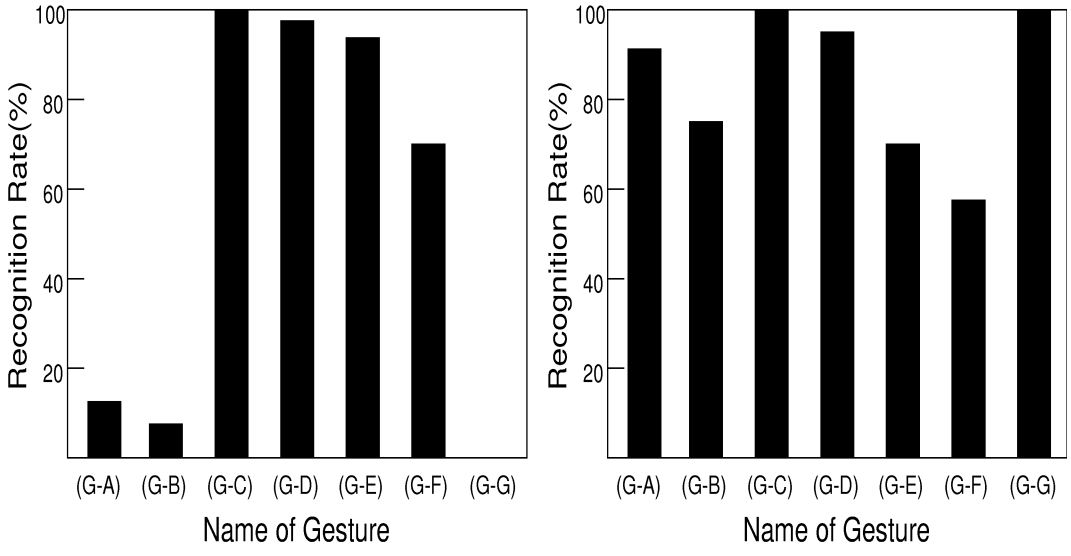
衣服Ｆ（黒のジャージ）

図 2.15　6 種類の衣服



（条件A）　　（条件B）

図 2.16　衣服（A,B,C,D,E,F）における平均認識率

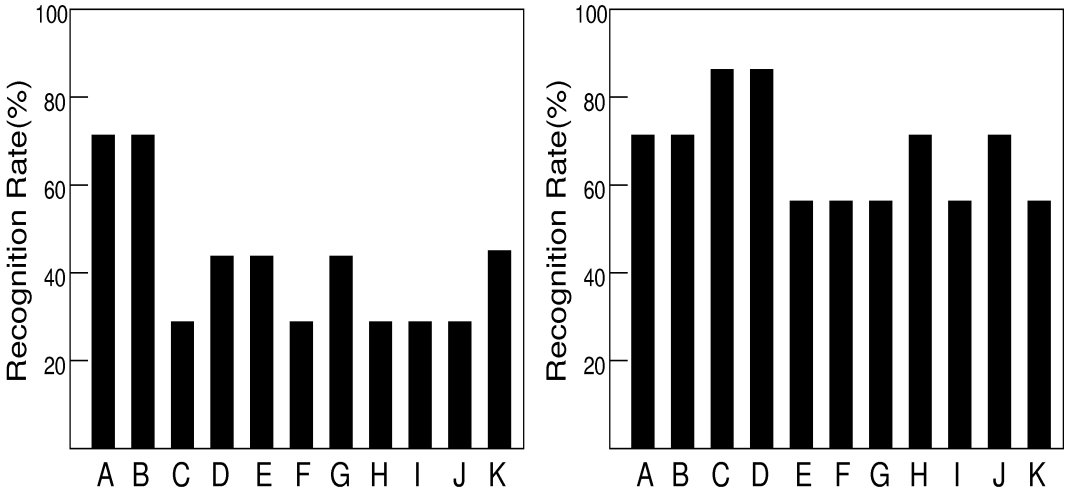


（条件A）　　（条件B）

図 2.17　身振り（G-A,B,C,D,E,F,G）における平均認識率

2.7.3 個人依存性評価実験

本節では同一被験者により作成された学習データを用いて他の被験者の身振りの認識実験を行った際の実験結果を報告する．衣服 A の標準身振りサンプルを用いて身振り標準パターンを登録し，（条件Ａ）衣服 A の標準身振りサンプルのみの場合と（条件Ｂ）衣服（A,B,C,D,E,F）のすべての類似身振りサンプルの場合の２通りの条件の下でプロトコル学習させた後，１１名の被験者の身振りサンプル（学習段階で用いていない未知かつ他人のテストサンプルで各被験者につき７種類の身振り動作，合計７７サンプル）の認識実験を行った．図 2.19 に１１名の被験者のスナップショットを示す（各写真は「鳥のまね」のスナップショットである）．各被験者（A–K）における平均認識率を図 2.18 に示す.



（条件Ａ）　　（条件Ｂ）

図 2.18　他者身振りサンプルの認識結果

プロトコル学習の際に（条件Ｂ）の類似身振りサンプルを用いることにより，図 2.18 における各被験者の平均認識率が全体で平均 27(%)（41(%) から 68(%) に）改善されている．この結果は，類似身振りサンプルをプロトコル学習させることにより，個人に比較的依存しない認識処理が可能になることを示している.

A

B

C

D

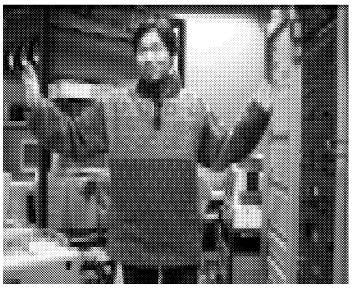
E

F

G

H

I



J

K

図 2.19　１１名の被験者

さらに，本実験で利用したすべての被験者１１名のテストサンプルは例えば「あなたが思い浮かべる“バイバイ”の身振りをしてください」のように自由発想の身振り動作を求めて収集したにも関わらず，全サンプルでの平均認識率が 68(%) に達したことは，各個人が行う身振り動作には共通性が内在していることを示唆している．

2.7.4 身振り情報の推定実験

「バイバイ」の動作を反復した際の身振り情報の推定結果を，身振り位相値については図 2.20 に，身振り速度は図 2.21 に，身振り振幅については図 2.22 にそれぞれ示す．これらの結果は，提案手法により身振り情報を実時間でフレーム毎に推定することが可能であることを示している．また，身振り情報を活用することにより，仮想物体の身振りによる操作が可能となる．

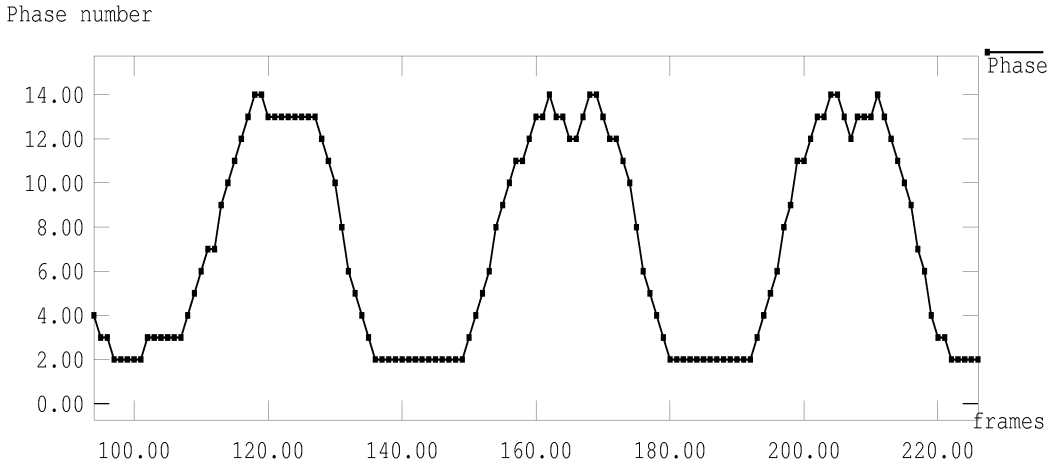


図 2.20 「バイバイ」の身振り位相値の変化

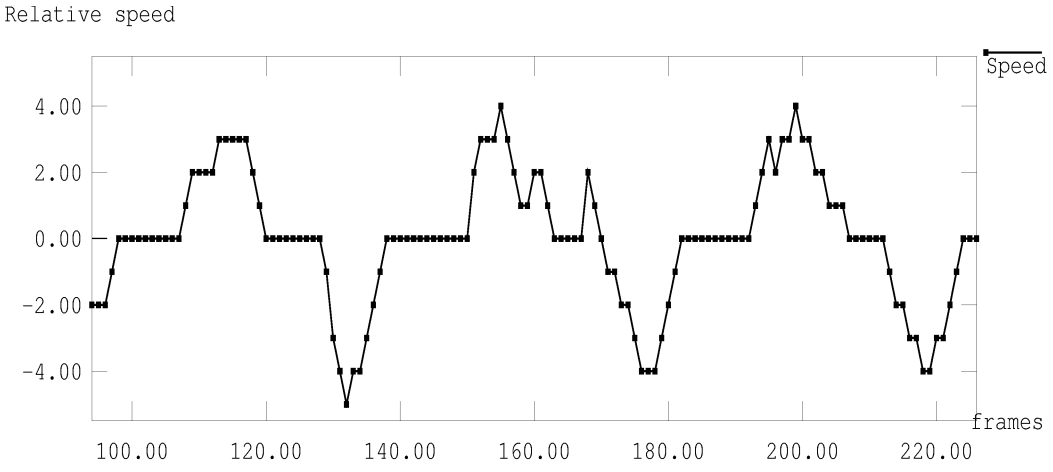


図 2.21 「バイバイ」の身振り速度の変化

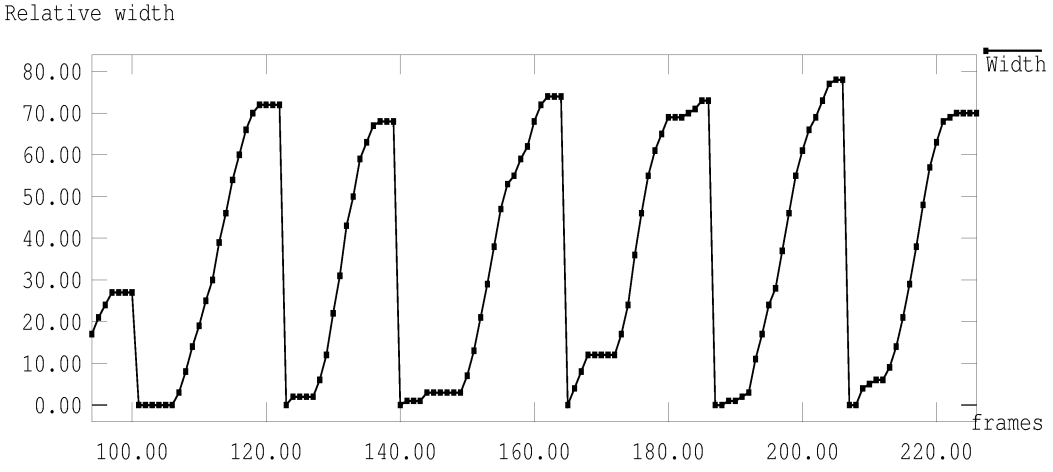


図 2.22 「バイバイ」の身振り振幅の変化

2.7.5 プロトコル学習の収束性評価実験

本節では「バイバイ」の身振り動作を「右手を振っても左手を振っても良い」と仮定して行った評価実験の結果を示す．実験手順は次の通りである．

［１］右手を一度だけ左右に振ることにより，身振り標準パターンを登録する．

［２］右手と左手を交互に振ることにより上記の身振りプロトコルを学習させる．

［３］右手を振った際と左手を振った際の評価値の差を逐次記録する．

［４］学習効率の指標として評価値の差に加えて式 (2.12) により身振りクラス重みを算出し記録する．

実験の結果得られた評価値の差の変化と身振りクラス重みの変化の様子を図 2.23 に示す．図 2.23 には，左手を振ることにより身振り標準パターンを登録して同様の実験を行った際の結果も重ねて示した．

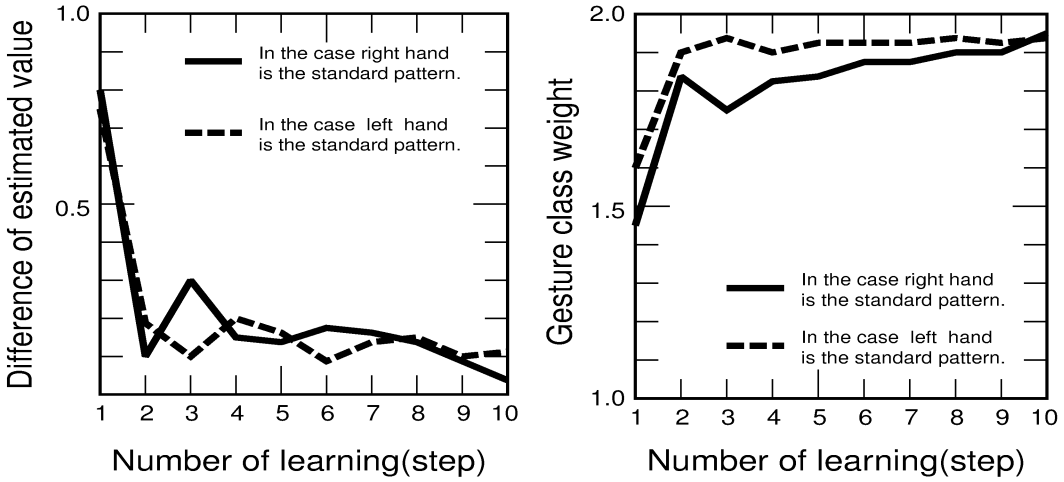


図 2.23 プロトコル学習の収束性

図 2.23 より，数回の学習のみで右手と左手を振った際の評価値の差は急激に減少した後に安定していることが分かる．身振りクラス重み W_i に関しても数回の学習のみで急激に増加した後に安定している．また，身振り標準パターンを右

手で登録した場合でも左手で登録した場合でも同様な学習曲線が得られている．以上より，提案手法では数回の身振り教示のみで十分であることが分かった．こうした性質は，利用者の手間をかけずコンピュータに身振りを学習させる際，非常に有効である．

2.7.6 認識処理の実時間性評価実験

提案手法の実時間性を評価するため，身振り標準パターンを徐々に増加させた際の処理速度の変化を調べた．図 2.24 にシステム性能の変化を表すグラフを示す．ここでのシステム性能とは 1 秒間に処理可能な身振り画像の枚数とする．図 2.24 からも分かるように，提案手法はソフトウェア処理のみで実時間での認識処理を達成している．

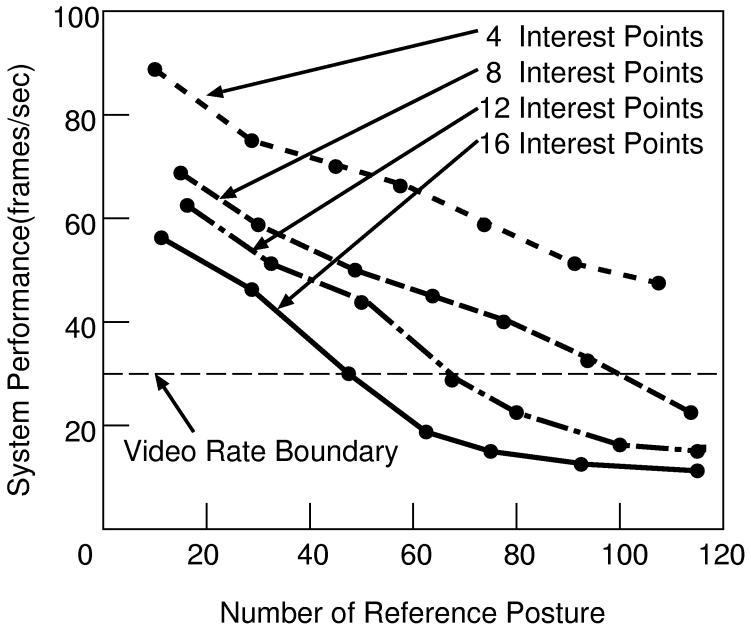


図 2.24 システム性能の変化

3 多注視点選択制御法

多注視点身振り認識法において n 種類の注視点に基づく認識処理を行う場合，単一の注視点に基づく認識処理の n 倍の照合コストを必要としてしまう問題がある．提案手法における照合処理は各注視点ごとに独立していることから，処理の並列化により照合コストの問題は解決できる．しかしながら，現在普及しているコンピュータの大半はシングルプロセッサにより動作しているため，これらの計算機環境においても認識性能を維持・向上させることへのニーズは大きい．

本章では，任意速度での身振り認識処理を実現するために，多注視点選択制御法を提案する．提案手法により以下の問題へと対応可能となることが期待できる．

［１］標準パターンの増加に伴う認識処理速度低下の問題，

［２］仮想現実感システムとの接続による認識処理速度低下の問題，

［３］ＯＳ環境下の他プロセスによる認識処理速度の不安定化の問題

なお，提案手法は多注視点身振り認識法を拡張することにより実現する．拡張後の多注視点身振り認識法の枠組みを図 3.1 に示す．

提案手法は，以下の３種類の選択制御系により構成される．

［１］パターン走査間隔の選択制御（Control 1），

［２］多注視点の選択制御（Control 2）．

［３］パターン照合間隔の選択制御（Control 3），

各選択制御は，図 3.1 中の Control 1 〜 Control 3 に対応している．

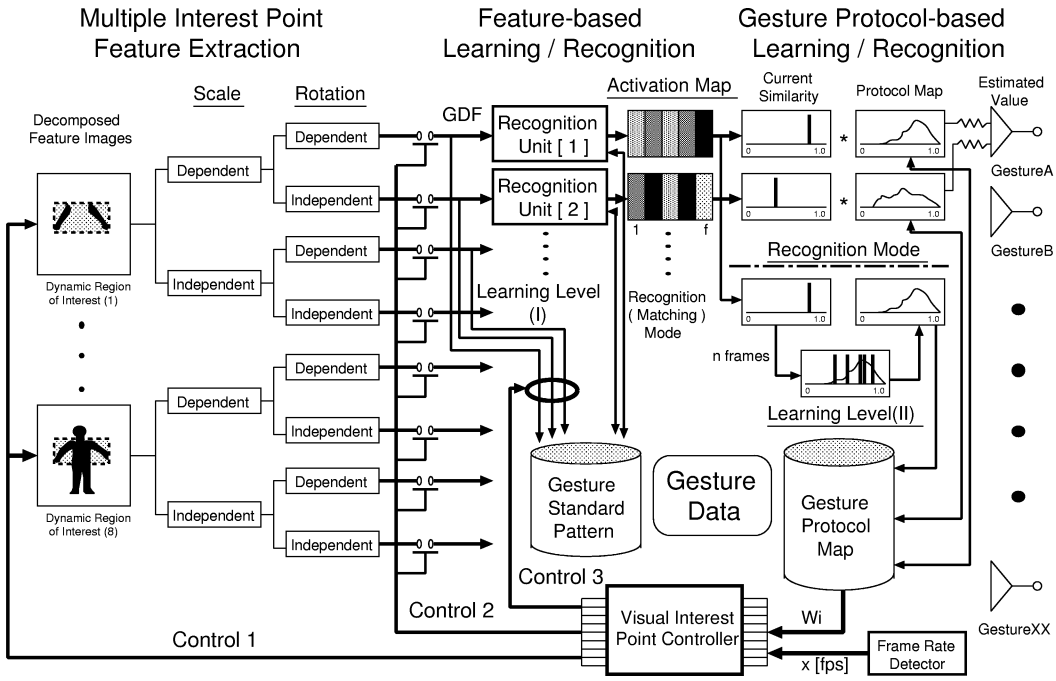


図 3.1 拡張後の多注視点身振り認識法の枠組み

3.1 選択制御手法の概要

身振り認識は適切な特徴の選択のみならず，観測座標系や特徴抽出の際の解像度選択の問題とも密接な関係があり，能動視覚問題 [42, 43, 44, 45, 46, 47, 48] の一つと捉えることができる．図 3.2 に示すようなフィードバック系を構成することにより，本手法では任意速度での身振り認識処理を実現する．

本論文では，多注視点選択制御を操作量 3 変数，制御量 1 変数（制御量はフレームレート x[fps]（目標値 v[fps])）のフィードバック制御問題として捉える．操作量はパターン走査間隔 S_k，パターン照合間隔 RS_k，有効注視点数 N_{vip} の 3 変数である．図 3.2 に示すように，フレームレート検出器からフレームレート情報を受け取った注視点コントローラ（Visual Interest Point Controller) では，まず微小操作量を適用した場合の処理時間の微小変化を検出してから，次の操作量を逐次決定して行く方式を採用し，制御偏差 e（$= v - x$）が最小となるように負のフィードバック制御を行う．

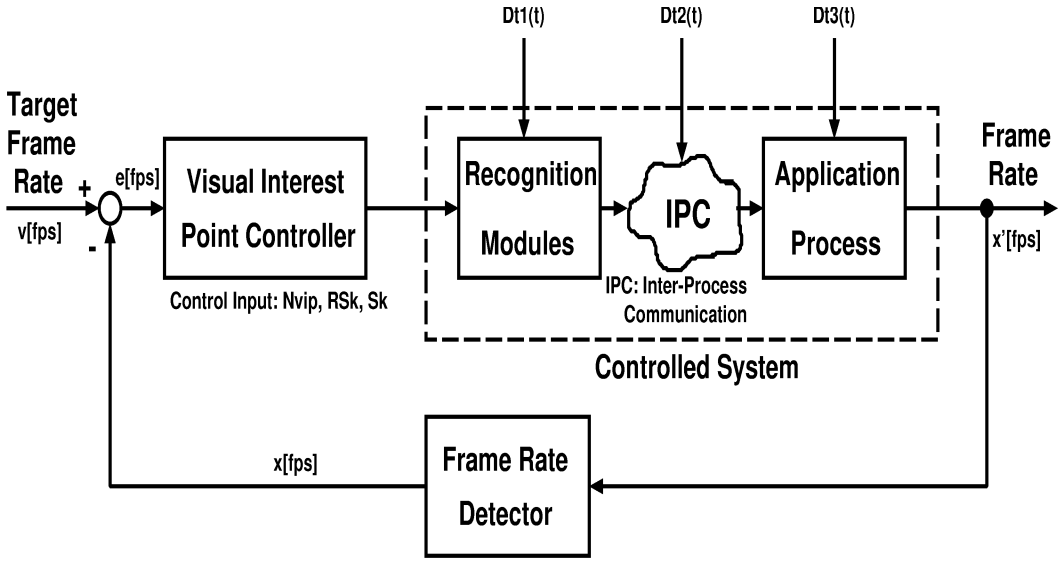


図 3.2 フィードバック制御のブロック図

本手法における制御指標（Control Index）S を式 (3.1) により定義する．

$$S = 3N_{MAX}(3 - S_k) + N_{MAX}(3 - RS_k) + N_{vip} \tag{3.1}$$

ここで N_{MAX} は最大注視点数，S は制御指標である．制御指標 S は大きいほど「処理速度は低下するがより多くの注視点を考慮した認識処理」が行われ，小さいほど「考慮される注視点の数は減少するが高速な認識処理」が行われることを表す．

図 3.3 に注視点コントローラの流れ図を示す．注視点コントローラでは，有効注視点の選択と特徴画像生成処理のオンオフ設定，パターン照合間隔とパターン走査間隔を設定する．想定可能な制御対象は，認識モジュール自体の負荷 $D_{t1}(t)$，アプリケーションとのプロセス間通信に伴うネットワーク負荷 $D_{t2}(t)$，さらにアプリケーション自体の負荷 $D_{t3}(t)$ の 3 種類であるが，本論文では認識モジュールの負荷 $D_{t1}(t)$ を制御することにより，制御対象全体での処理時間の総和を目標値へと安定化させることを試みる．

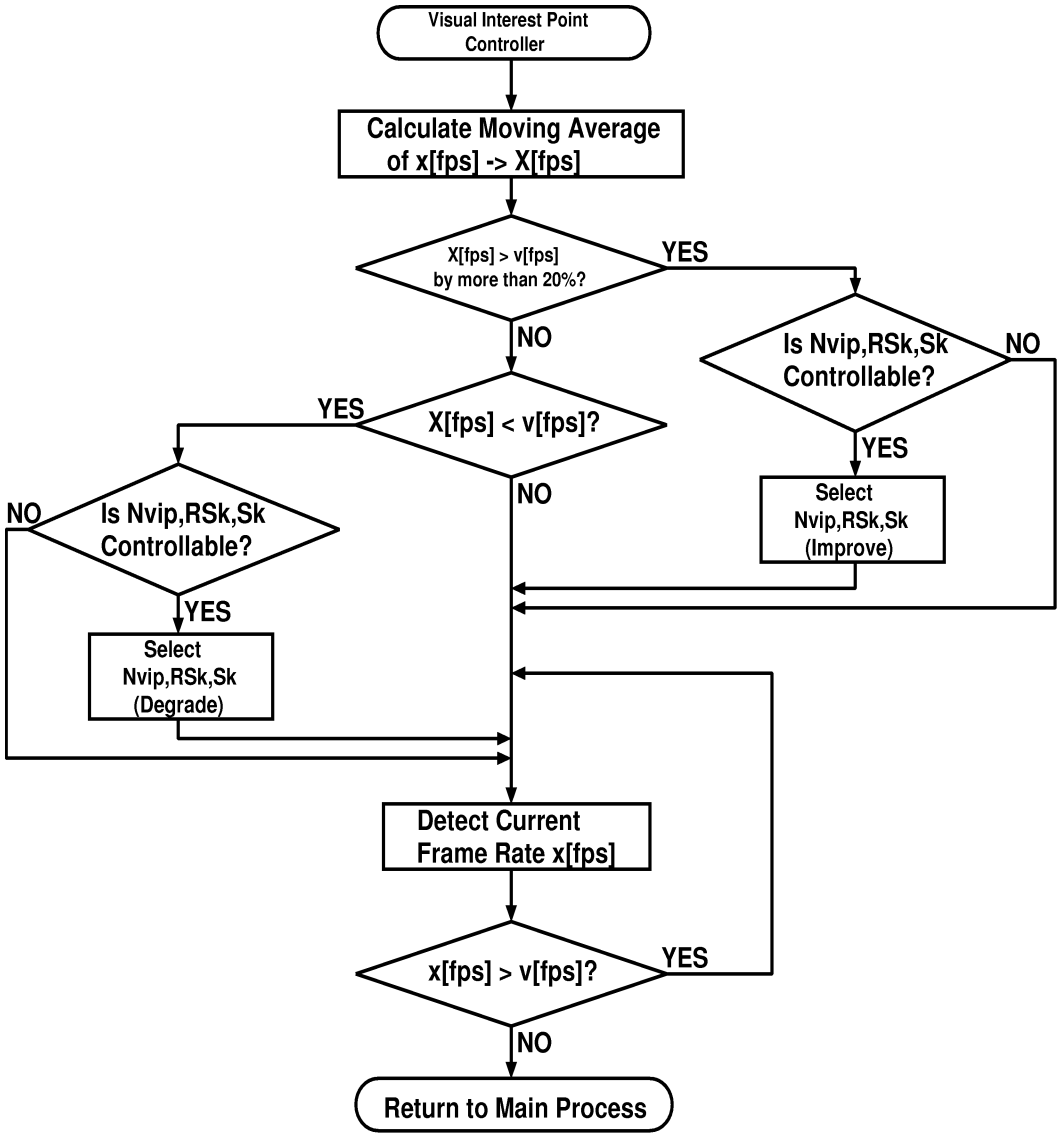


図 3.3 注視点コントローラの流れ図

3.2 パターン走査間隔の選択制御

多注視点身振り認識法では，複数種類の特徴画像を組み合わせることにより動的注視領域を設定している．このため，動的注視領域は利用可能な特徴画像の領域および図形パターンに依存する．本節では，新たにエッジ画像を加え，手話動作などのより広範な身振りの認識への対応を図る．

図 3.4 に示すように，差分画像とシルエット画像，さらにエッジ画像を利用して人物の位置・姿勢・動作に関する特徴量をフレーム毎に抽出する．差分画像からは，動的注視領域１（動作部位の形状とその重心位置），動的注視領域２（人物中心座標系から見た動作部位の相対的な位置とその形状），さらに動的注視領域５（画面内の動作部位の位置とその形状）についての特徴量を抽出する．シルエット画像からは，動的注視領域４（人物位置とその形状）や動的注視領域３（人物シルエットの形状），さらに動的注視領域８（動作領域内のシルエット形状）についての特徴量を抽出する．エッジ画像からは，動的注視領域６（人物輪郭の形状），さらに動的注視領域７（動作領域内のエッジ形状）についての特徴量を抽出する．以上の８種類の動的注視領域はそれぞれが常に独立しているのではなく，実際には対象の動作に依存して重なり合う．これらの特徴量を各動的注視領域について求めることにより，表 3.1 〜表 3.2 に示す３２通りの注視点に対応した形状特徴パターンを抽出する．なお，各動的注視領域からの形状特徴パターンの抽出方法については，2.2節で提案した手法に従う．

多様な注視点を得るためには，より多くの注視領域が必要となるが，注視領域は増やせば増やすほど，特徴抽出処理のコストを増大させてしまう．要求される処理速度を達成できない場合には，認識精度を犠牲にしてその要求を満たすことが必要となる．ここでは式 (2.3) の計算量を削減することによりこの要求を満たすことを試みる．パターン形状の把握で重要な角度分解能の低下は避け，半径方向パターンの走査間隔の増加により計算量を抑制する．具体的には，r の刻み間隔であるパターン走査間隔 S_k を変更するが，S_k の増加による認識精度の極端な低下を避けるために，刻み間隔は最大３（経験的数値）までに制限する．このため S_k の値域は $[1,3]$ に限定される．パターン走査間隔の選択制御は認識精度に直接的な影響を及ぼすため，他の操作量による制御での目標達成が困難となったときのみ実行する．

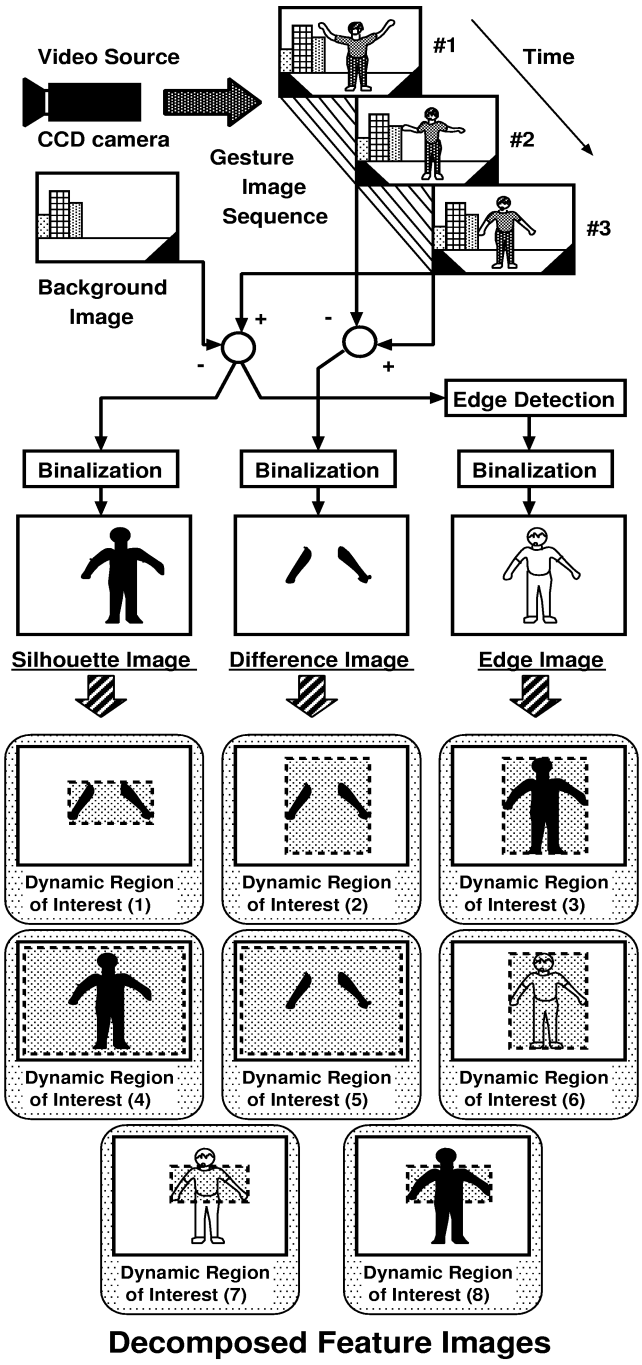


図 3.4 身振り画像の領域分割手法

表 3.1 設定した３２通りの注視点（その１）

注視点番号	位置	大きさ	回転	特徴情報源
1	非依存	依存	依存	動的注視領域 1
2	非依存	依存	非依存	動的注視領域 1
3	非依存	非依存	依存	動的注視領域 1
4	非依存	非依存	非依存	動的注視領域 1
5	依存	依存	依存	動的注視領域 2
6	依存	依存	非依存	動的注視領域 2
7	依存	非依存	依存	動的注視領域 2
8	依存	非依存	非依存	動的注視領域 2
9	非依存	依存	依存	動的注視領域 3
10	非依存	依存	非依存	動的注視領域 3
11	非依存	非依存	依存	動的注視領域 3
12	非依存	非依存	非依存	動的注視領域 3
13	依存	依存	依存	動的注視領域 4
14	依存	依存	非依存	動的注視領域 4
15	依存	非依存	依存	動的注視領域 4
16	依存	非依存	非依存	動的注視領域 4

表 3.2 設定した３２通りの注視点（その２）

注視点番号	位置	大きさ	回転	特徴情報源
17	依存	依存	依存	動的注視領域 5
18	依存	依存	非依存	動的注視領域 5
19	依存	非依存	依存	動的注視領域 5
20	依存	非依存	非依存	動的注視領域 5
21	非依存	依存	依存	動的注視領域 6
22	非依存	依存	非依存	動的注視領域 6
23	非依存	非依存	依存	動的注視領域 6
24	非依存	非依存	非依存	動的注視領域 6
25	依存	依存	依存	動的注視領域 7
26	依存	依存	非依存	動的注視領域 7
27	依存	非依存	依存	動的注視領域 7
28	依存	非依存	非依存	動的注視領域 7
29	依存	依存	依存	動的注視領域 8
30	依存	依存	非依存	動的注視領域 8
31	依存	非依存	依存	動的注視領域 8
32	依存	非依存	非依存	動的注視領域 8

3.3 パターン照合間隔の選択制御

パターン照合処理において要求される処理速度を達成できない場合には，前節で述べたパターン走査間隔の選択制御に加えてパターン照合間隔の選択制御によりこの要求を満たすことを試みる．具体的には，式 (2.4) による形状特徴パターン照合の際の身振り標準パターン番号を任意の間隔だけスキップさせることにより計算量の抑制を図る．しかしながら，パターン照合間隔 RS_k の増加により照合対象とならなかった身振り標準パターンの位相値は考慮されないため，位相値推定の精度が低下する問題がある．位相値推定の精度は主に有効注視点数に依存するため，有効注視点数を多く設定できる場合はパターン照合間隔 RS_k の増加に伴う精度低下を補うことができる．

しかしながら，有効注視点数が少ない場合 RS_k 低下の影響は顕在化し，位相値情報を利用するアプリケーションシステムに悪影響を与えるため，パターン照合間隔 RS_k は最大３（経験的数値）までに制限する．このため RS_k の値域は $[1, 3]$ に限定される．RS_k を増やす際には有効注視点数 N_{vip} を増加させることにより，位相値推定の精度低下を可能な限り回避する．

3.4 多注視点の選択制御

認識処理全体において目標速度に達しない場合，有効注視点数を削減する多注視点の選択制御を行う．多注視点の選択制御では，３２種類の注視点のオンオフ処理を行えるため，選択制御の自由度を十分に確保できるメリットがある．具体的には式 (2.8) により得られる注視点重みを参照してオンオフ処理を行う．この際，注視点重みの小さい順に削減することにより，それぞれの注視点の重要度を反映させた上での処理速度の改善を行う．有効注視点数の削減の結果，不要となる特徴画像については生成処理を停止する．

多注視点の選択制御における操作量は有効注視点数 N_{vip}（値域は $[1, 32]$）であり，N_{vip} が大きいほど速度は犠牲となるが頑健さの高い認識処理が実現され，反対に，N_{vip} が小さいほど頑健さは犠牲となるが高速な認識処理が実現される．

以上の提案手法における計算量は（パターン走査量）×（パターン照合量）×

（有効注視点数）に比例する．従って，パターン走査間隔の選択制御とパターン照合間隔の選択制御および多注視点の選択制御を組み合わせた場合の最小計算量は，最大計算量の $\frac{1}{3} \times \frac{1}{3} \times \frac{1}{32} = \frac{1}{288}$ 倍となり，きめ細かく負荷を制御することができる．

3.5 認識系の再構成手法

本節では認識系の再構成手法について説明する．本手法の目的は，個別の注視点をオフすることにより不要となる処理は停止させ，可能な限り認識処理速度の向上に貢献することにある．

入力画像から始まり各注視点に対応する特徴量の抽出に至るまでの処理構成を図 3.5 に示す．図 3.5 からも分かるように，提案手法における初期視覚関連の処理は表 3.3 に示すような４つのパターン処理の階層から構成されている．

表 3.3 処理の階層

階層	ノード数	処理対象
1	1	入力画像
2	3	特徴画像
3	8	動的注視領域における図形パターン
4	３２	各注視点に対応する形状特徴パターン

パターン処理のオンオフは以下の基本方針に従って行う．

［１］階層４の注視点がすべてオフになっている動的注視領域の処理は停止する．

［２］注視点のオフに伴って不要となる形状特徴パターンの正規化処理は停止する．

［３］動的注視領域がすべてオフになった特徴画像の生成処理は停止する．

以上の方針を具体化するための手法を次に示す．

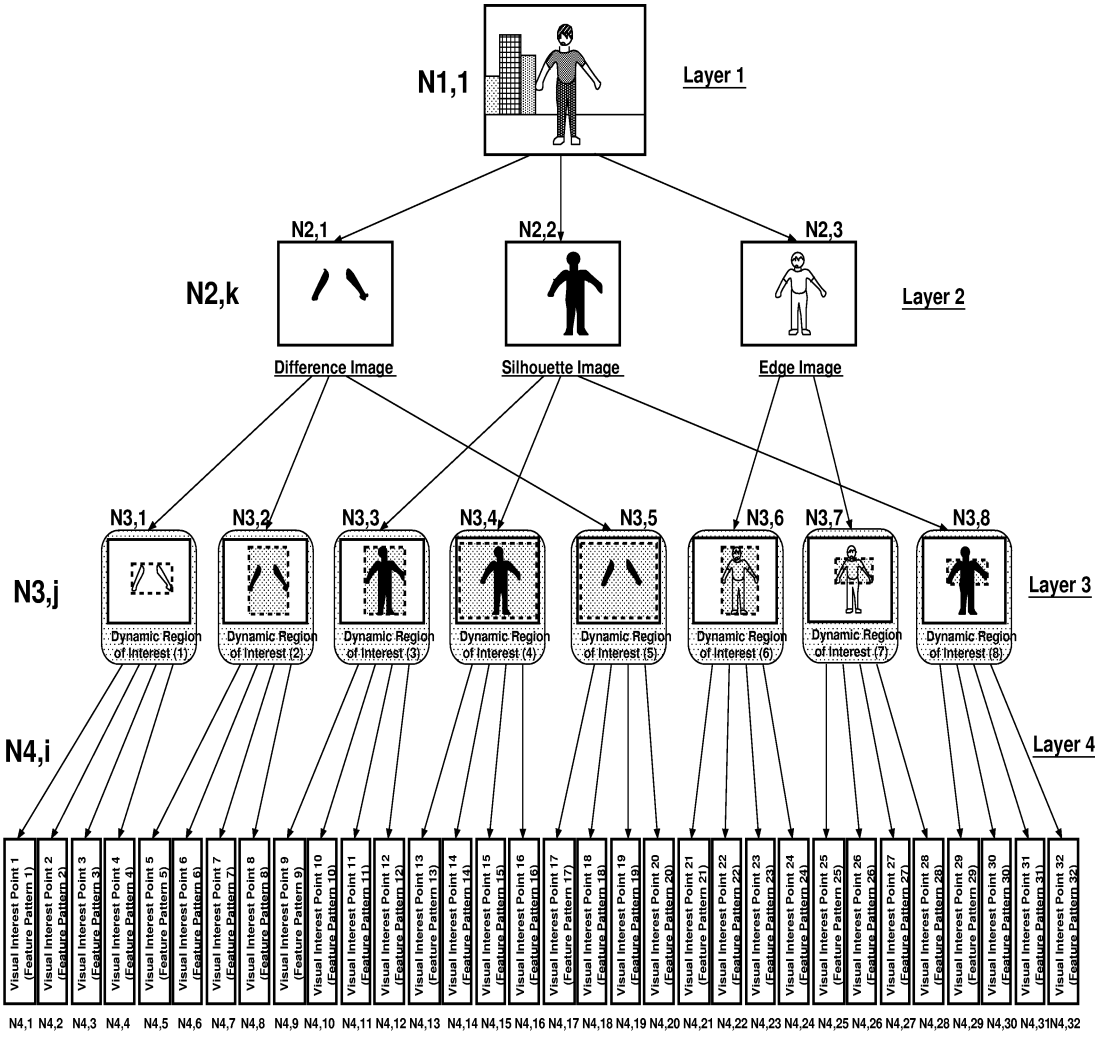


図 3.5 認識系の内部構成図

階層４における任意ノード $N_{4,i}$（$1 \leq i \leq 32$）に，注視点がオンの時 “1”，オフの時 “0” の値を与え，階層３における任意ノード $N_{3,j}$（$1 \leq j \leq 8$）の値を，

$$N_{3,j} = \sum_{i:withN_{3,j}} N_{4,i:withN_{3,j}} \tag{3.2}$$

により定義する．同様に，階層２における任意ノード $N_{2,k}$（$1 \leq k \leq 3$）の値を，

$$N_{2,k} = \sum_{j:withN_{2,k}} N_{3,j:withN_{2,k}} \tag{3.3}$$

により定義する．さらに，階層１におけるノード $N_{1,1}$ の値を，

$$N_{1,1} = \sum_{k:withN_{1,1}} N_{2,k:withN_{1,1}} \tag{3.4}$$

により定義する．

階層４におけるノード数は３２であるので注視点がすべてオンの場合，最上層のノード $N_{1,1}$ の値は３２となる．なお，すべてのノードが０に設定されることのないように，最低１つの注視点がオンになるように制限する．各ノードの値は，そこでの処理に対する必要性の程度を表す．

以上の手法により各ノードの値を求め，その値が０となった場合には，そこのノードにおける処理を停止させる．本手法により最下位ノードのオンオフ設定のみで，上位ノードでの処理を下位ノードの状態と連動させながら処理コストの削減を図ることが可能になる．

3.6 評価実験

提案手法による身振りインタフェースシステムをパーソナルコンピュータ（Intel Pentium MMX 266MHz，OS:Linux）上にC言語で実装し，評価実験を行った．ＣＣＤカメラにより撮影された画像は画像入力装置（IBM Smart Capture Card I）を通してパーソナルコンピュータに解像度横 80[dot] 縦 60[dot] の大きさで取り込まれ，オンライン認識される．なお，すべての処理をソフトウェアで行っている．また，特別な照明や背景は使用せず通常の室内で行った．

本実験では，勾配係数 $a = 5.0$，位相項 $\phi = 0$，カーネル数 $\Omega = 1$，注視点強調係数 $\alpha = 0.1$ （経験的数値），分離係数 $\beta = 500$（経験的数値），解像度 $\Theta = 64$（経験的数値），最大注視点数 $N_{MAX} = 32$，初期有効注視点数 $N_{vip} = 32$，初期パターン走査間隔 $S_k = 1$，初期パターン照合間隔 $RS_k = 1$，初期制御指標 $S = 288$ の条件下で評価した．

3.6.1 応用システムとの接続実験

本節では仮想現実感アプリケーションを実際に接続したときの提案手法の有効性を検証する．アプリケーション例として，ジェスチャービデオシステムを考案・開発した．本研究におけるジェスチャービデオシステムは，利用者の動作に合わせて任意のビデオ画像をリアルタイムで再生・提示し，仮想オブジェクトとのインタラクションを可能とする． 従来のインタラクティブなビデオシステムの多くは，CD-ROM 検索などによる動画像再生機能の提供に留まっているが，提案手法に基づく身振りインタフェースシステムにより，利用者とビデオ画像との直接的なインタラクションを実現できる．この際，十分なインタラクションのリアリティを利用者に与えるにはシステムの動作速度を少なくともビデオレート程度に維持する必要がある [34, 35]．図 3.6 に身振りインタフェースシステムとアプリケーションシステムを接続する際のネゴシエーション手順を示し， 図 3.7 にジェスチャービデオシステムのブロック図を示す．

図 3.7 に示すように提案システムは，身振りインタフェースシステム，ジェスチャービデオシステム，画像編集システムの大別して３つの要素から構成される．

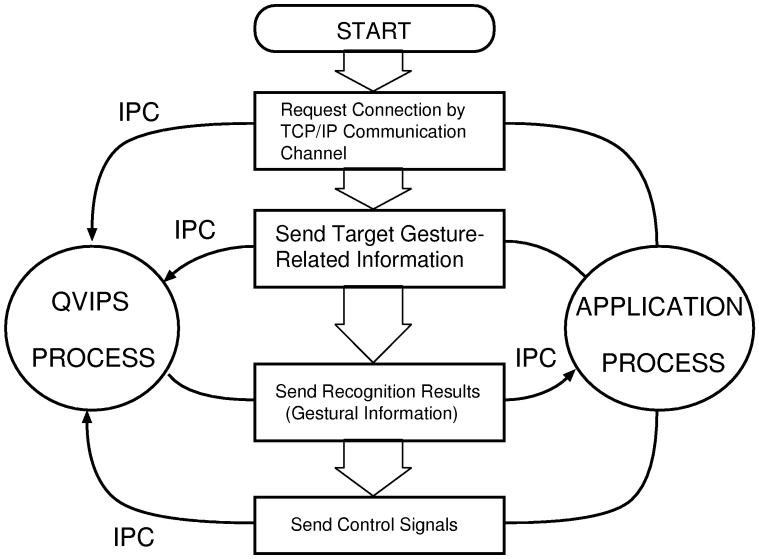


図 3.6 接続手順

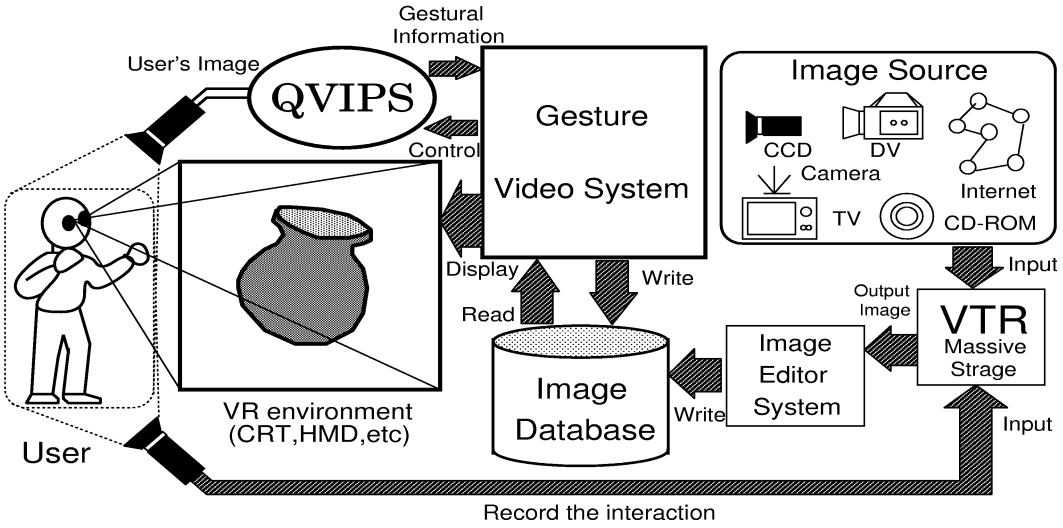


図 3.7 ジェスチャービデオシステムのブロック図

図 3.6 に示すように身振りインタフェースシステムとアプリケーションシステム間では，TCP/IP プロトコルによるプロセス間通信により身振り情報およびシステム制御信号の伝達が双方向に行われる．なお，ジェスチャービデオシステムにおける主な負荷は画像表示に関する処理である．本システムの実行画面例を図 3.8 に示す.

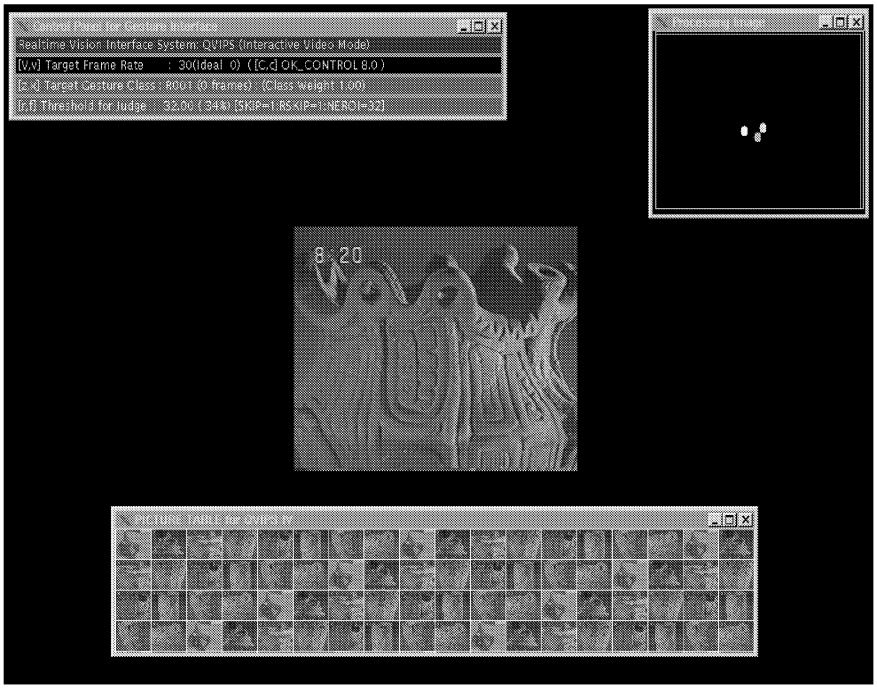


図 3.8 ジェスチャービデオシステムの実行画面例

接続実験は次の手順に従って行う.

［１］図 3.6 に示す接続手順により，身振りインタフェースシステムとジェスチャービデオシステムとのコネクションを確立する.

［２］身振りインタフェースシステム側の目標フレームレート v[fps] を 30[fps] に設定する.

［３］身振り動作の例として「物体を眺める（体を左右に揺らす）動作」（図 3.9 を参照）をプロトコル学習させる.

［４］数回身振り動作を繰り返した後，多注視点選択制御を開始する．この際，身振りインタフェースシステム側でフレームレート x[fps] と制御指標 S を記録しておく．なお，本実験では利用者の身振り動作と同期して，任意角度から見た埴輪の画像が表示される．

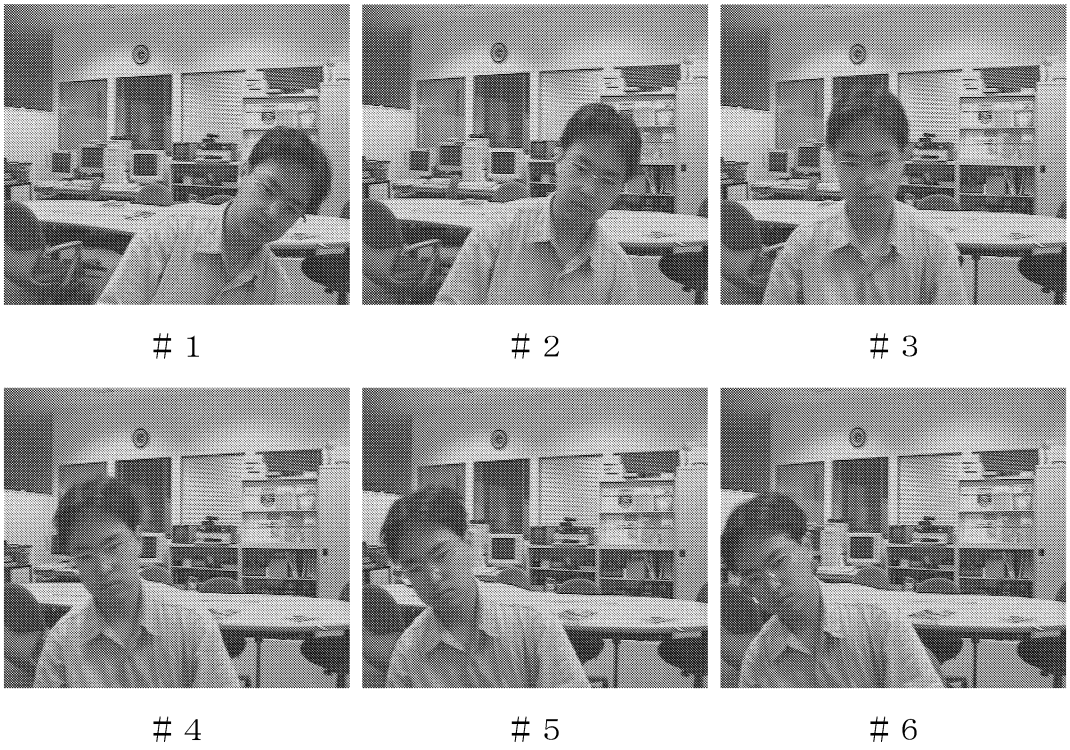


図 3.9 物体を眺める動作のスナップショット

以上の手順を経て得られたシステムフレームレートの応答を図 3.10 に示し，制御指標 S の応答については図 3.11 に示す． 多注視点選択制御が開始される 5120 フレーム以降では，システム動作速度は目標値である 30[fps] へと徐々に近づき，5260 フレーム以降は目標値に安定していることが分かる．以上の結果は，実際のアプリケーションシステムとの接続により発生する，処理速度の低下と不安定化の問題を，提案手法により克服できることを示している．

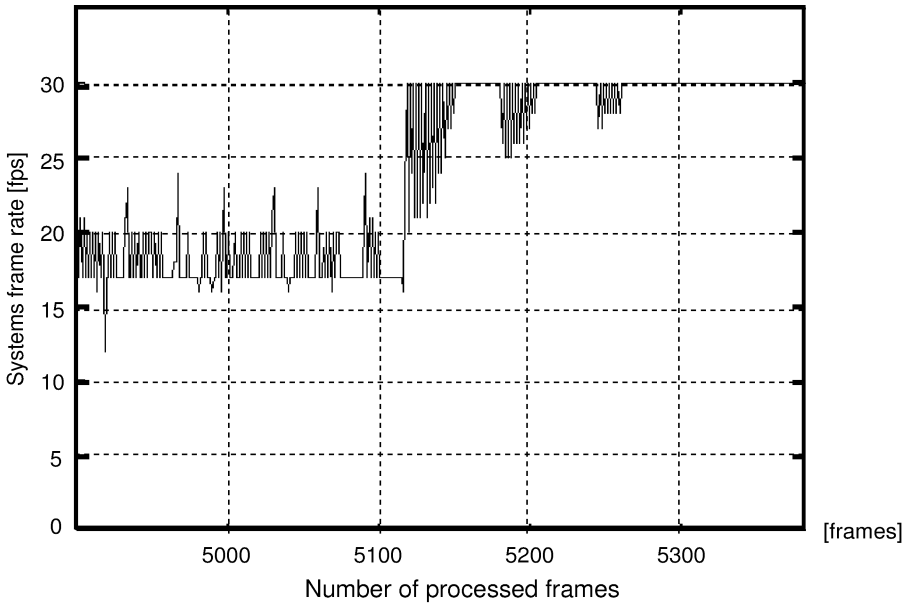


図 3.10 システムフレームレートの応答

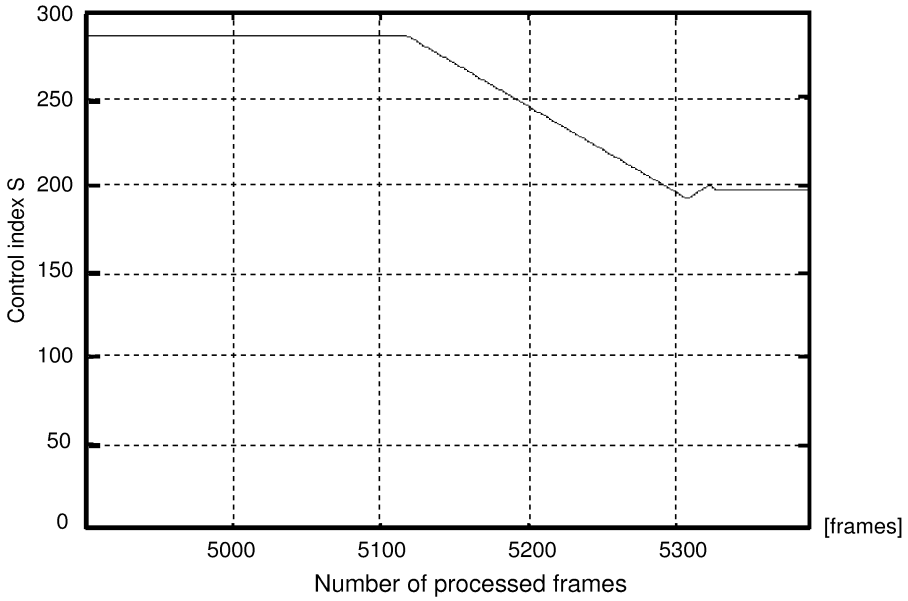


図 3.11 制御指標 S の応答

3.6.2 認識率への影響評価実験

多注視点選択制御の身振り認識率への影響を調べるために，１０通りのフレームレート条件（1 〜 45[fps]（5[fps] 間隔））下で，表 3.4 に示すグループＡの手話動作，表 3.5 に示すグループＢの手話動作，表 3.6 に示すグループＣの手話動作を認識させる実験を行った．なお，本実験により得られる結果の信頼性を向上させるため，３つのグループの手話動作について評価実験を行った．図 3.12 にグループＡのＧＡからＧＨまでの８種類の手話動作，図 3.13 にグループＢのＧＡからＧＨまでの８種類の手話動作，図 3.14 にグループＣのＧＡからＧＨまでの８種類の手話動作の軌跡画像（動作の向きを表す矢印付き）を示す．

表 3.4 グループＡの手話動作

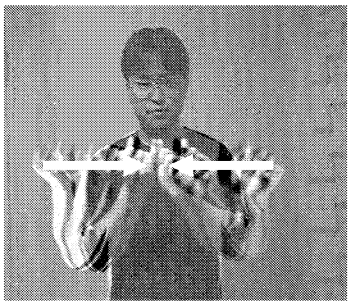
手話動作名 [49]	省略表現	標準画像列の枚数
「結婚」	GA	26
「ダメ」	GB	20
「ＯＫ」	GC	16
「おめでとう」	GD	19
「バイバイ」	GE	13
「ご無沙汰してます」	GF	21
「生まれる」	GG	21
「姉」	GH	22
合計		158

表 3.5　グループＢの手話動作

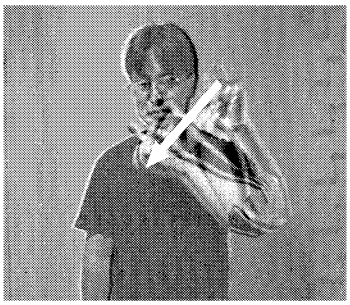
手話動作名 [49]	省略表現	標準画像列の枚数
「体育」	GA	18
「海」	GB	33
「暇」	GC	22
「忙しい」	GD	25
「平等」	GE	25
「値上げ」	GF	15
「情報」	GG	14
「コミュニケーション」	GH	27
合計		179

表 3.6　グループＣの手話動作

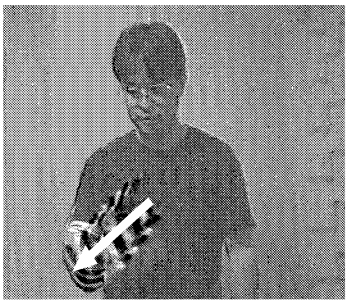
手話動作名 [49]	省略表現	標準画像列の枚数
「注目させる」	GA	14
「頭にくる」	GB	18
「驚く」	GC	23
「一生懸命」	GD	14
「思う」	GE	11
「わがまま」	GF	19
「平気」	GG	18
「つまり」	GH	25
合計		142



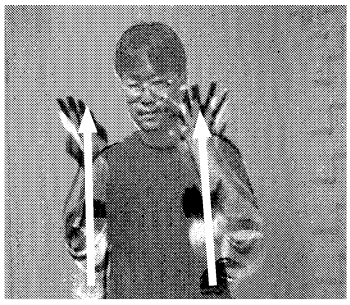
ＧＡ



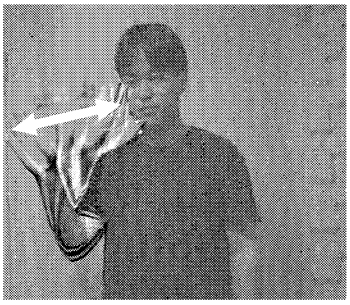
ＧＢ



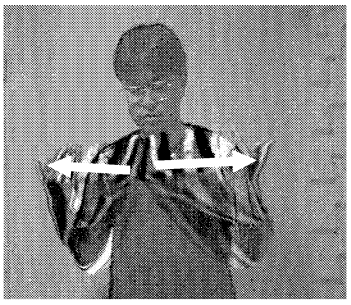
ＧＣ



ＧＤ

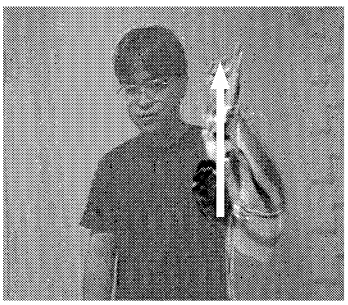


ＧＥ



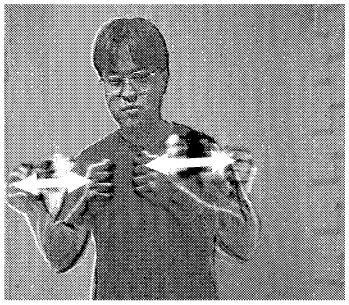
ＧＦ

ＧＧ

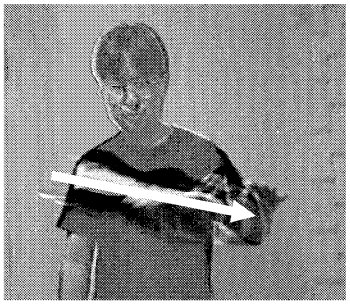


ＧＨ

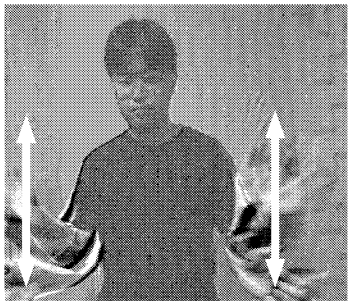
図 3.12　グループＡの手話動作



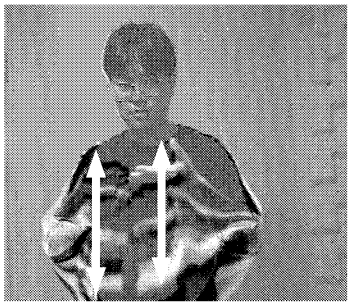
ＧＡ



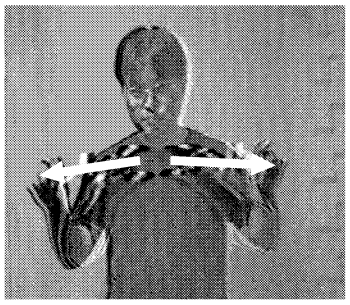
ＧＢ



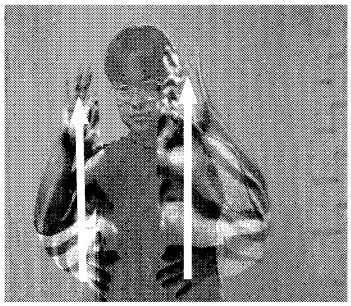
ＧＣ



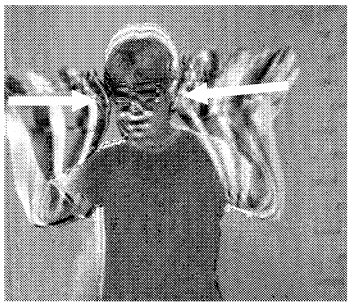
ＧＤ



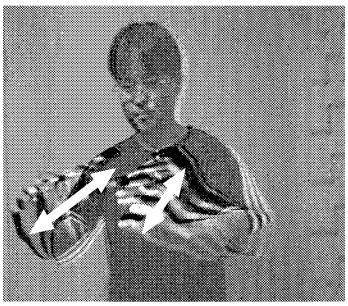
ＧＥ



ＧＦ



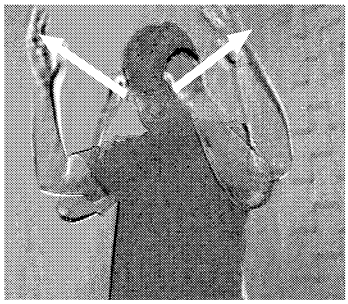
ＧＧ



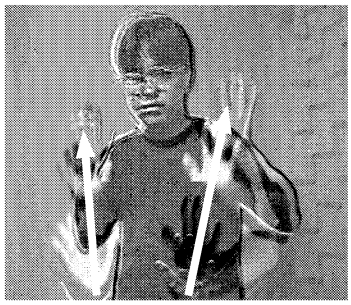
ＧＨ

図 3.13　グループＢの手話動作

ＧＡ

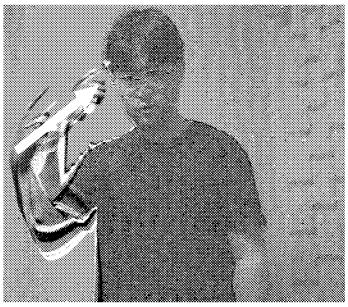


ＧＢ

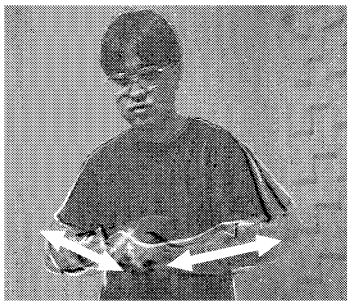


ＧＣ

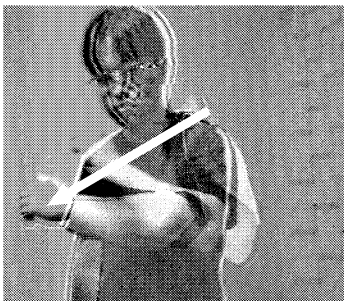
ＧＤ



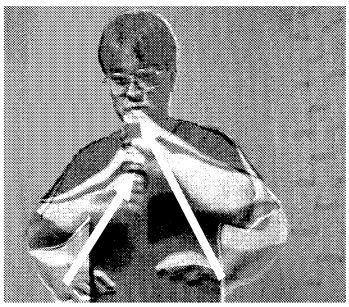
ＧＥ



ＧＦ



ＧＧ



ＧＨ

図 3.14　グループＣの手話動作

評価実験は，各グループの手話動作に対して次の手順に従い行った．

［１］８種類の手話動作画像の身振り標準パターンを登録する，

［２］８種類の手話動作画像についてプロトコル学習させる，

［３］目標フレームレートを設定する（1 〜 45[fps]（5[fps] 間隔）），

［４］多注視点選択制御を開始する，

［５］テストサンプル（各グループの各手話動作につき８個のサンプル）を認識させる．

以上の実験手順で得られた結果を，各フレームレート条件下での平均認識率については図 3.15 に，各手話動作の平均認識率については図 3.16 に示す．さらに，各フレームレート条件下での制御指標 S を図 3.17 に示す．

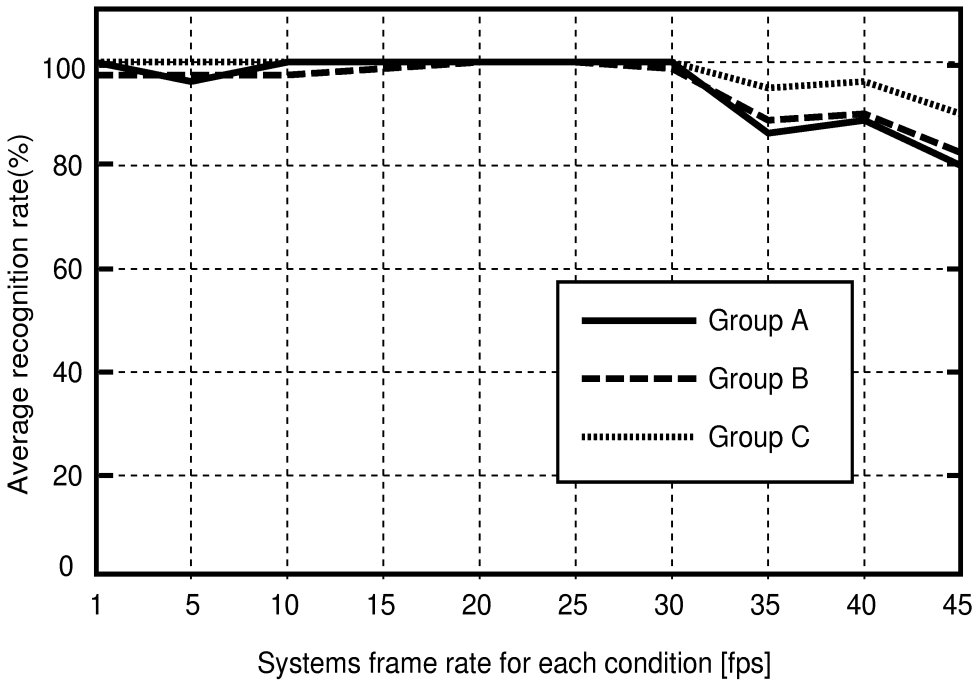


図 3.15 各フレームレート条件下での平均認識率

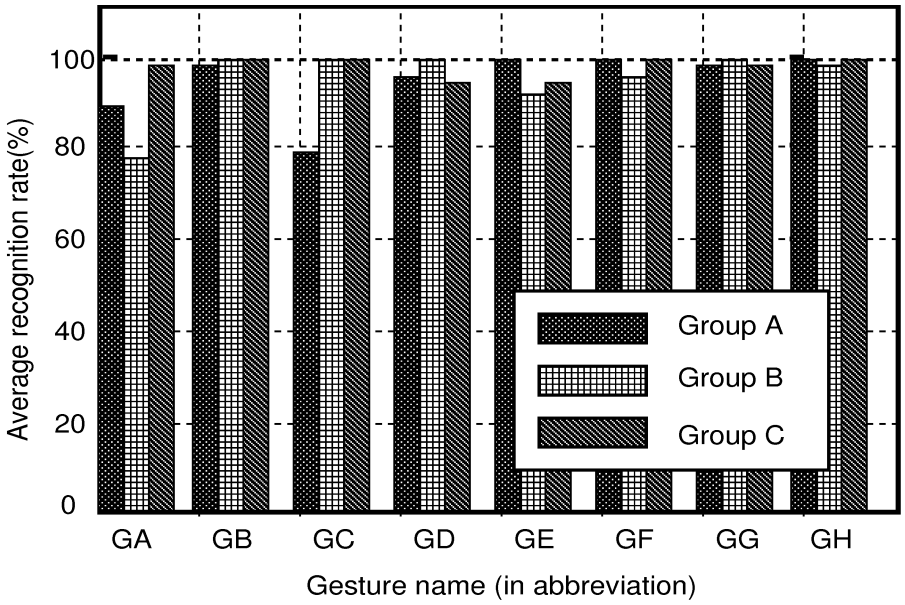


図 3.16 各手話動作の平均認識率

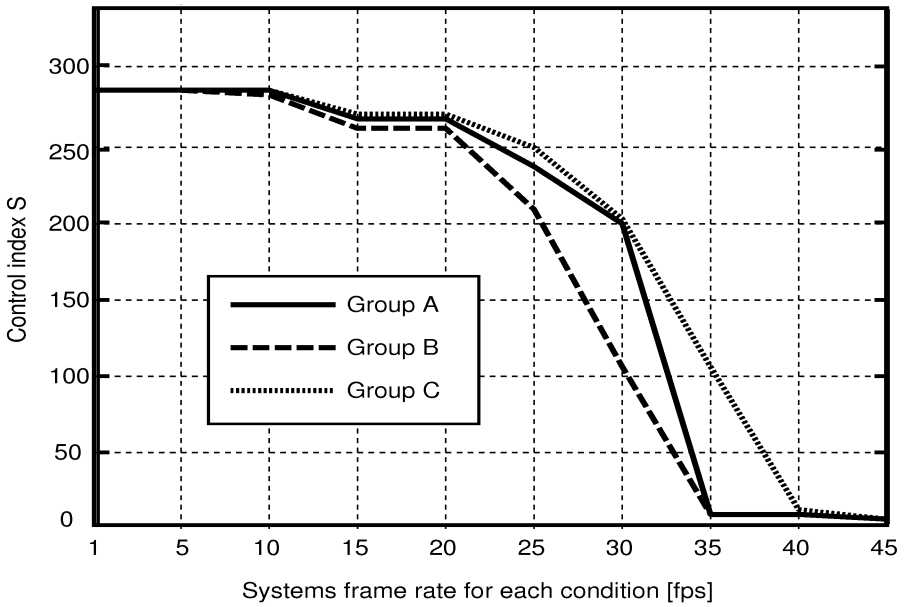


図 3.17 各フレームレート条件下での制御指標 S

ここでの認識率は，認識処理時の有効画像数に占める（正答数）を（誤答数＋正答数）で割ることにより算出した．なお，多注視点選択制御を行わない場合のシステムの動作速度は，グループＡの実験において 10[fps]，グループＢの実験において 9[fps]，グループＣの実験において 11[fps] であった．

図 3.15 からも分かるように，すべてのグループにおいて 1[fps] ～ 30[fps] の領域では，平均認識率の低下はほとんど見られない．この結果は，今回選定した動作群については多注視点身振り認識法がフレームレートの影響を受けずに高い認識性能を発揮したことを示している．また，提案手法によりビデオレート（30[fps]）を大幅に上回る動作速度が実現されることが分かった．

一方，35[fps] を超える領域では平均認識率の有意な低下が見られるが，この理由としては以下の２点が挙げられる．

［１］実験に使用したデジタルビデオは NTSC 規格に準拠しているため，30[fps] を越える領域では同一の画像が再度入力されてしまうことがある．同一の画像が入力されると画像差分量が検出されなくなるため，認識処理は停止する．この結果，本来継続すべきところの処理が中断され，正確な評価値が算出されなくなる．この問題は，NTSC 規格以外の高速画像記録再生装置を使用することで解決できる．

［２］実験に使用した画像入力装置は 30[fps] 以上の速度で画像を入力できるが，40[fps] 辺りからは，画像データが不完全な状態のままで読み出されてしまうことがある．これは，画像入力装置の欠陥である可能性が高いが，基本的にこの装置も NTSC 規格に準拠しており，想定されている以上の速度で読み出すこと自体に問題があると言える．

上述の理由から，35[fps] を越える領域に見られる平均認識率の有意な低下は，提案手法によるものではなく，周辺の画像入出力機器がその原因になっていると考えられる．

いずれにしても１０通りのフレームレート条件（1 ～ 45[fps]（5[fps] 間隔））下で，グループＡについては平均 95(%)，グループＢについては平均 94(%)，グループＣについては平均 98(%) の認識率を得た．この結果は，提案手法により認識率の大幅な低下を招かずに任意速度での認識処理が実現されることを示している．

一方，図 3.16 に示した結果からも分かるように，各手話動作の平均認識率については一部に 80(%) を若干下回る結果があるが，その他については 90(%) 以上の平均認識率が得られている．この結果は，提案手法により手話動作の種類に依存しない認識処理が実現されることを示している．

さらに，図 3.17 に示した結果からも分かるように，制御指標 S は 15[fps] を越えた辺りから徐々に低下し始め，30[fps] を超えたところで急激に低下している．この結果は，30[fps] を越える辺りから，認識処理に要する時間と画像入力に要する時間が拮抗し始め，認識処理側での負荷を急速に減じる必要性が生じたことを示している．

図 3.17 において，３つのグループの中で標準画像列の最も多いグループＢにおける制御指標 S の曲線は，他のグループの場合よりも早く下降し始めている．逆に，最も標準画像列の少ないグループＣにおける制御指標 S の曲線は，他のグループの場合よりも緩やかに下降していることが分かる．以上の結果は，各フレームレート条件下での制御指標 S の曲線の勾配が登録される標準画像列の枚数に依存することを示している．

3.6.3 認識処理の実時間性評価実験

提案手法の有効性を調べるために，身振り標準パターンを徐々に増加させた際のシステムフレームレートの変化を，提案手法を適用する場合と適用しない場合で計測した．なお，提案手法を適用する際には，身振りインタフェースシステム側の目標フレームレート v[fps] を 30[fps] に設定して計測した．

図 3.18 から分かるように，いずれの場合でもソフトウェア処理のみで実時間認識可能であるが，提案手法を適用しない場合（FRate1）は，身振り標準パターンの増加にフレームレートが反比例することが分かる．一方，提案手法を適用した場合（FRate2）には，身振り標準パターンが増加しても常時 30[fps] のフレームレートが実現されている．

以上の結果は，標準パターンの増加に伴う認識処理速度低下の問題が提案手法により克服可能であることを示している．

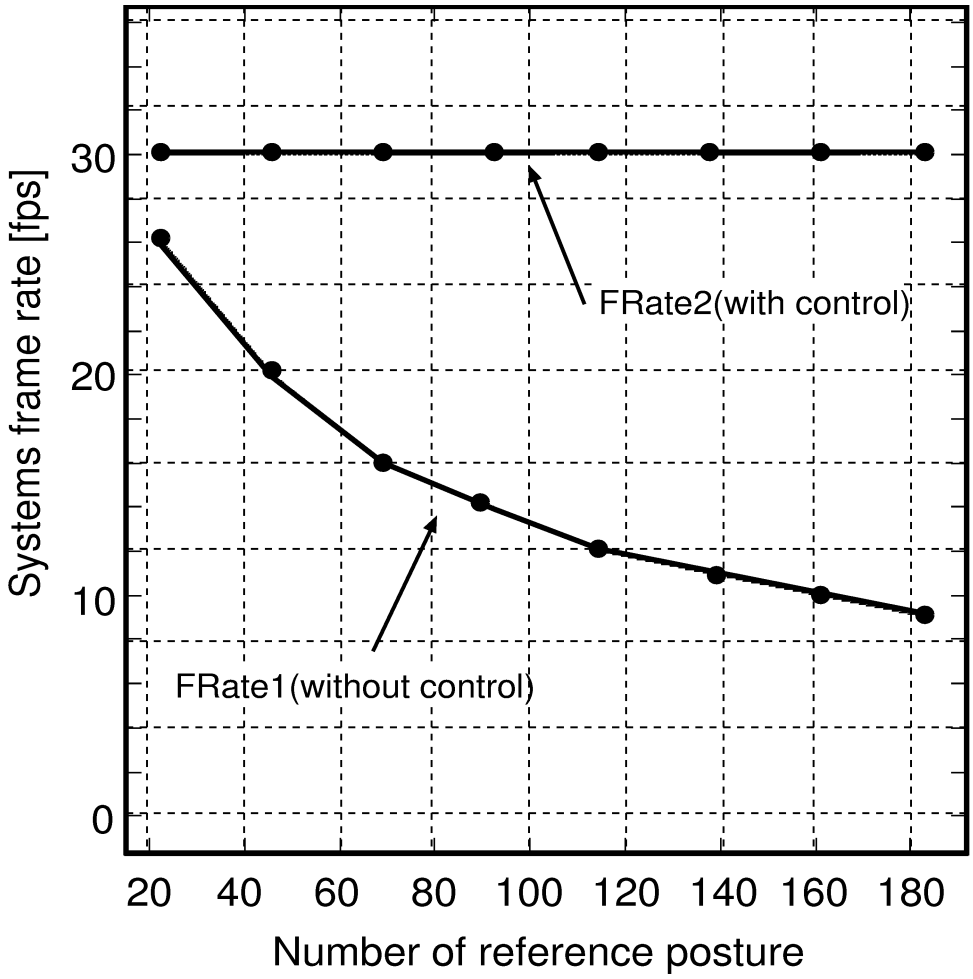


図 3.18 システムフレームレートの変化

4 手話画像データベース検索への応用

身振りインタフェースに画像検索機能を付与することにより，仮想物体とのインタラクションのみならず画像データベース検索を同時に行うことが可能になる．身振りによる画像データベースへの直接的なアクセスは，身振りを媒介とした人間－機械系インタラクションの有効性を高める可能性が高い．

本章では，身振りにより直接検索することが従来困難であった手話画像データベースに焦点を当てる．手話画像データベースの検索は同一人物の多様な動作を対象とするため，色情報を主要な手がかりとする従来手法 [50] では対応が困難である．これまでに本論文では，プロトコル学習が手話動作を含めた身振り画像の認識に有効であることを示してきた．具体的には，複数の類似動作をプロトコル学習させることにより，視覚的な共通項を重視した認識処理が実現されることを評価実験により示した．本章では，提案手法を活用した手話画像データベース検索システムの構成方法を示し，評価実験によりその有効性を明らかにする．

なお，手話画像においては，話者の視線方向や表情変化といった微妙な動きについても考慮する必要があるが，ここでは比較的大雑把な動作による類似手話動作検索を対象とするに留める．

4.1 検索システムの構成方法

検索システムは，多注視点身振り認識法と多注視点選択制御法により実装された身振りインタフェースシステムに，データベース検索マネージャ機能を追加することにより構成する．図 4.1 にデータベース検索マネージャの流れ図を示し，図 4.2 に手話画像データベース検索システムのブロック図を示す．

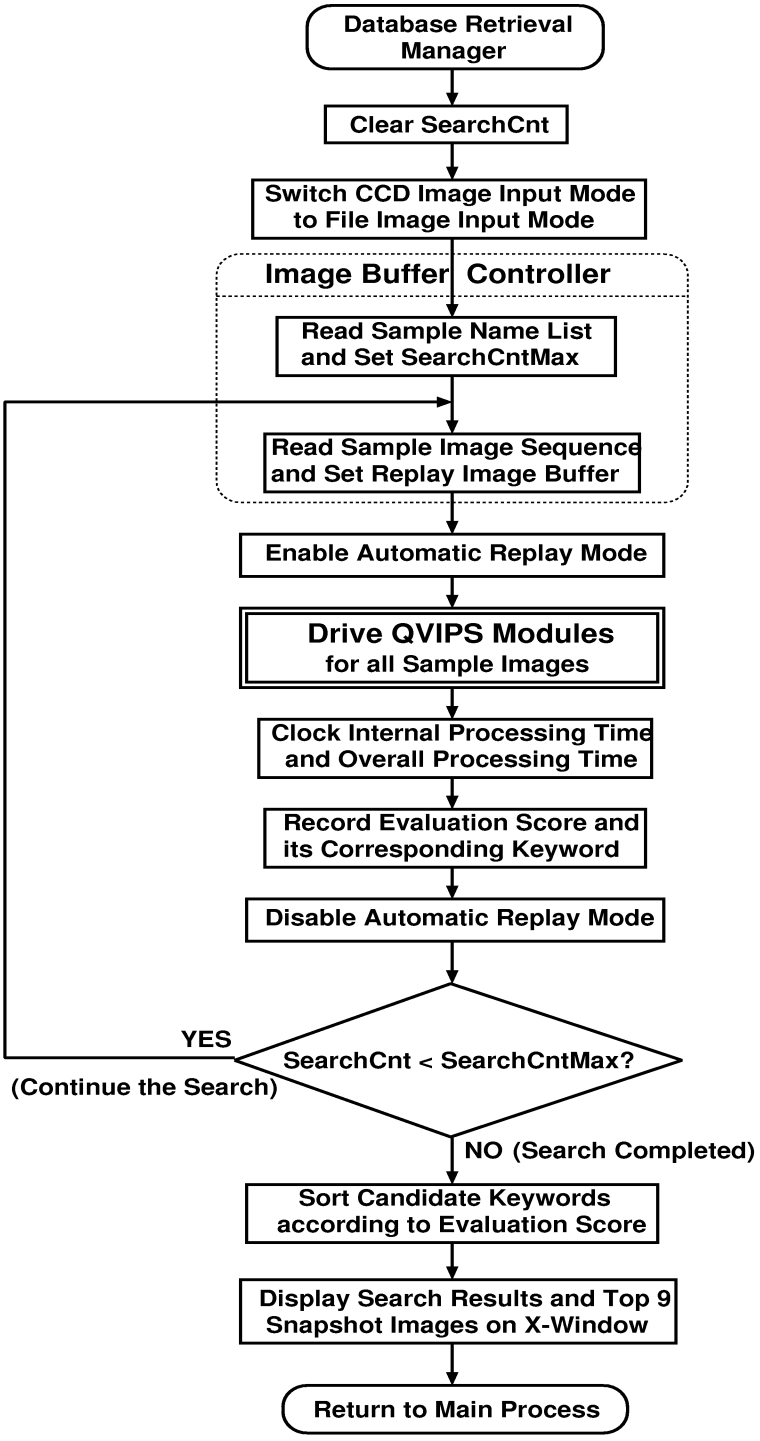


図 4.1 データベース検索マネージャの流れ図

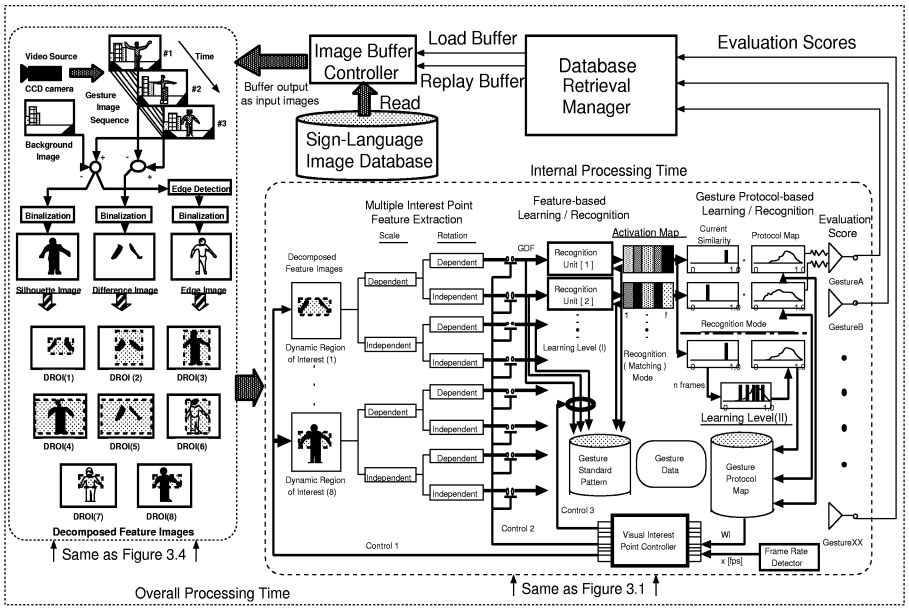


図 4.2 手話画像データベース検索システムのブロック図

データベース検索マネージャでは次の手続きを実行する.

［１］画像入力モードを外部入力モードからファイル入力モードに切り替える.

［２］画像バッファコントローラにより手話画像データベースから画像サンプルを読み出した後，認識処理系へと手話画像を入力する.

［３］システム全体での処理時間（Overall Processing Time）と内部処理時間（Internal Processing Time）を計測する.

［４］認識処理の結果として得られる評価値（多注視点身振り認識法による評価値 $E^{(i)}$）と該当動作名を保存する．データベース内のすべての手話画像サンプルの評価が終了するまで［２］～［４］を繰り返す.

［５］検索結果を評価値の降順にソートし，先頭９候補のスナップショット画像と動作名をウィンドウ上に表示する.

手話画像の学習や評価処理のみならず，利用者インタフェースもこれまでに開発してきたシステムを利用できるため，QVIPS を駆動させるための制御部を新たに実装するだけで良い．なお，提案システムにおいては，プロトコル学習の結果に基づいて検索処理が実行される．

提案手法の原理上，学習サンプルには，あらかじめ登録されている画像ファイルのみならず，ＣＣＤカメラなどの撮像装置からの画像をそのまま利用することも可能である．具体的に提案システムでは，

［１］事前に作成しておいた画像ファイルを訓練サンプルとして使用するモード，

［２］ＣＣＤカメラからの入力画像を直接訓練サンプルとして使用するモード，

以上の２通りに対応している．このことは，利用者が手話動作を行い，それと類似する手話動作の動作名をその場で検索することも可能であることを意味する．こうした機能は従来の手話画像データベース検索システムには存在しなかったものである．なお，本実験では検索対象の手話画像サンプルを手話画像データベースから読み出すため，ＣＣＤカメラからの画像入力モードは利用せずに行う．

一方，多注視点選択制御法により，有効注視点の数を任意に設定することで，高速な検索処理のみならず，数多くの視覚的特徴を同時に考慮した柔軟な検索処理を行うことができる．あらかじめ時間を指定して検索する機能（以降，指定時間検索と呼ぶ）も，多注視点選択制御法により可能となる．指定時間検索では，データベース検索処理の最中においても動的に注視点の選択制御を行い，制限時間での検索を実現する．

4.2 評価実験

提案システムは，パーソナルコンピュータ，画像キャプチャーカード，ＣＣＤカメラにより構成される．提案システムの構成を図 4.3 に示す.

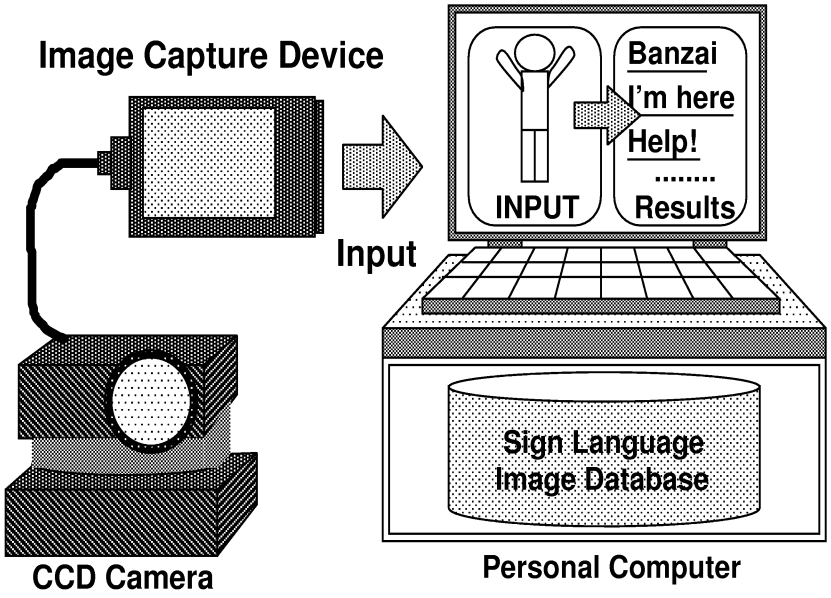


図 4.3 提案システムの構成

手話画像データベースとしては，文献 [49] を参考にして６４種類の手話動作を撮影し記録した． 手話画像データベースに登録されている画像の総数は８６４枚であり，１手話動作は平均１４枚の濃淡画像により構成されている．画像サイズは横 320[dot] 縦 240[dot]，フレームレートは 15[fps] にてサンプリングされる．収録した全手話動作のスナップショット画像とその軌跡画像を付録Ｃ（→手話画像データベース収録画像）に示す．なお，本実験で使用する画像データベースは，厳密な手話画像データベースを意図して作成されたものではなく，同一人物の多様な動作を検索するシステムの評価を主目的として作成された.

提案システムは，手話画像サンプルを多注視点身振り認識法により学習した後，データベース全体に渡り画像サンプルの評価処理を行い，最終的には候補動作名とその評価値を出力する．検索結果の表示画面例を図 4.4 に示す．なお，手話画

像サンプルは検索の際に横 80[dot] 縦 60[dot] の解像度に変換された後，評価処理される．

以上に述べたシステムを利用して，勾配係数 $a = 5.0$，カーネル数 $\Omega = 1$，注視点強調係数 $\alpha = 0.1$，分離係数 $\beta = 500$ と設定した後，評価実験を行った．

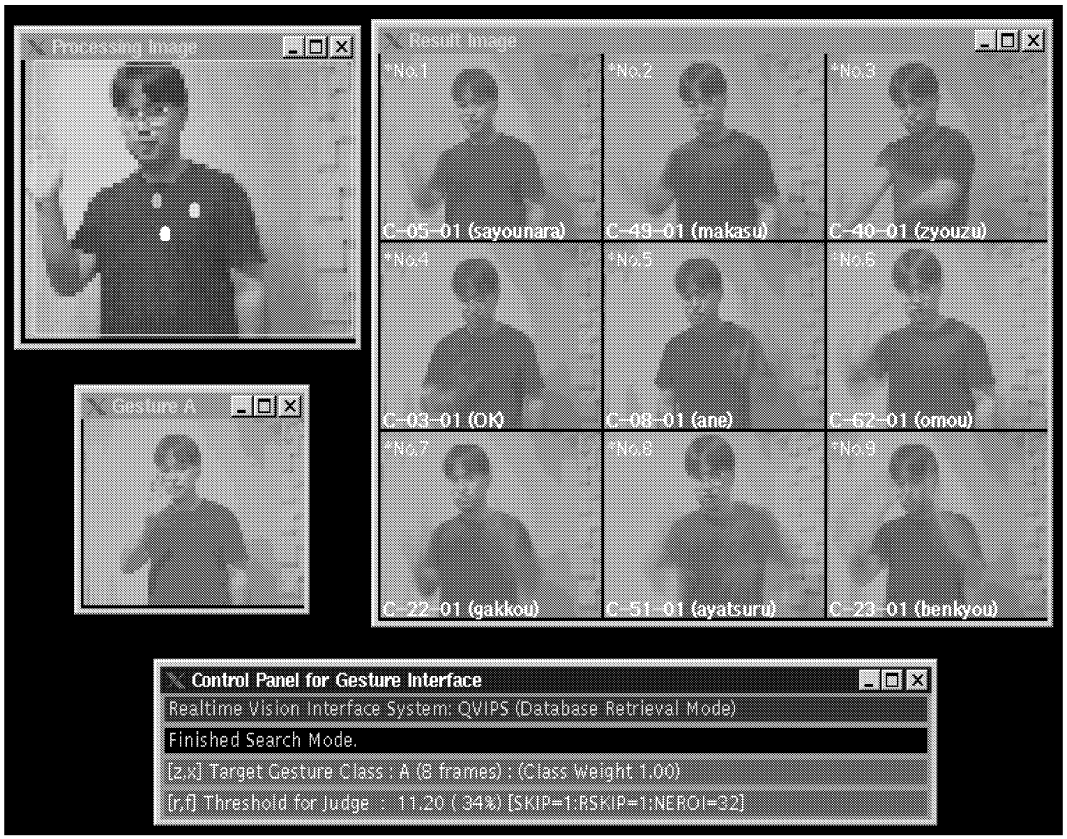


図 4.4 検索結果の表示画面例

4.2.1 類似手話動作の検索実験

手話画像データベースの中から任意で手話画像サンプルを選び，提案手法によりプロトコル学習させた後，類似手話動作の検索実験を行った．手話画像サンプルとして「バイバイ」,「情報」,「雲」,「操る（あやつる）」の４種類の手話動作を採用した．なお，本実験では，多注視点選択制御は行わないものとする．

検索結果における上位１０番までの候補とその評価値を表 4.2 と表 4.3 に示す．「バイバイ」の検索結果における各候補に対応する軌跡画像を図 4.5 ,「情報」の検索結果における各候補に対応する軌跡画像を図 4.6 ,「雲」の検索結果における各候補に対応する軌跡画像を図 4.7 ,「操る」の検索結果における各候補に対応する軌跡画像を図 4.8 にそれぞれ示す．

「バイバイ」の結果については，第１位から第６位まで，すべて片腕による手話動作が類似動作候補として挙げられている．「バイバイ」は片腕を左右に振る動作であり，これらの結果は妥当である．「情報」の結果については，第１位から第８位まで，すべて両腕による手話動作が類似動作候補として挙げられている．「情報」は広げた両腕を狭める動作であるが，第３位の「頭にくる」は狭めた両腕を広げる動作である．提案システムでは動作の向きは考慮していないので，この結果は妥当であると言える．「雲」の結果については，第１位から第４位まで，すべて両腕による手話動作が類似動作候補として挙げられている．「雲」は両腕を広げた状態で手のひらを繰り返し表裏に返す動作であり，第２位の「平等」は胸の前で両腕を広げる動作である．データベース内に「雲」と類似する動作が少ないことから，第２位から第４位まではすべて両腕を広げたり狭めたりする動作が候補として挙げられている．「操る」の結果については，すべて両腕による手話動作が類似動作候補として挙げられている．「操る」は，右腕と左腕を交互に胸の前で上下させる動作であり，第４位の「スポーツ」では腰から胸にかけての位置で，右腕と左腕を交互に前後させる動作である．提案システムでは，動作の奥行きは考慮していないので，これらは妥当な結果であると言える．

次に，各手話画像サンプルの画像枚数と検索所要時間，平均処理速度を表 4.1 に示す．表 4.1 に示すように，多注視点選択制御を行わない場合でも，ビデオレート（30[fps]）を大きく上回る速度で類似手話動作を検索できることが分かった．

表 4.1 各手話動作における検索性能

手話動作名	画像枚数	所要時間 [s]	平均処理速度 [fps]
バイバイ	10	15.16	57.16
情報	11	15.80	54.86
雲	15	17.21	50.38
操る	19	18.70	46.36

表 4.2 「バイバイ」と「情報」の検索結果

「バイバイ」					
No	動作名	評価値	No	動作名	評価値
1	バイバイ	36.00	6	思う	2.54
2	任す	3.98	7	学校	2.29
3	上手	3.16	8	操る	2.19
4	ＯＫ	3.02	9	勉強	2.19
5	姉	2.74	10	長い	2.18

「情報」					
No	動作名	評価値	No	動作名	評価値
1	情報	35.55	6	音楽	0.019
2	一生懸命	0.035	7	値上げ	0.013
3	頭にくる	0.034	8	暇	0.011
4	結婚	0.030	9	バイバイ	0.009
5	学校	0.024	10	水曜日	0.008

第１位　第２位　第３位　第４位　第５位

第６位　第７位　第８位　第９位　第１０位

図 4.5　「バイバイ」の検索結果

第１位　第２位　第３位　第４位　第５位

第６位　第７位　第８位　第９位　第１０位

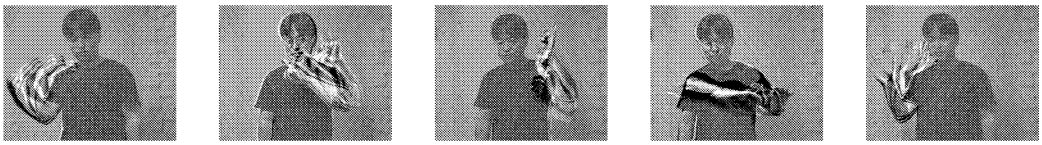
図 4.6　「情報」の検索結果

表 4.3 「雲」と「操る」の検索結果

「雲」					
No	動作名	評価値	No	動作名	評価値
1	雲	34.46	6	任す	2.9E-3
2	平等	9.8E-3	7	ダメ	2.8E-3
3	注目	6.9E-3	8	姉	1.6E-3
4	遠い	6.5E-3	9	海	1.5E-3
5	上手	3.1E-3	10	バイバイ	1.5E-3

「操る」					
No	動作名	評価値	No	動作名	評価値
1	操る	33.88	6	地震	9.21E-4
2	成長	2.95E-3	7	浮気	9.02E-4
3	暇	2.94E-3	8	長い	8.27E-4
4	スポーツ	2.16E-3	9	深い	3.94E-4
5	体育	1.70E-3	10	交流	3.41E-4

第１位　第２位　第３位　第４位　第５位



第６位　第７位　第８位　第９位　第１０位

図 4.7 「雲」の検索結果

第１位　第２位　第３位　第４位　第５位

第６位　第７位　第８位　第９位　第１０位

図 4.8 「操る」の検索結果

4.2.2 別クラスでの学習による検索実験

本節では複数の手話動作を別々の動作クラスとしてプロトコル学習させた後，検索処理を行って得られた結果とその考察を示す．本実験では「バイバイ」を動作クラスＡとして,「ダメ」を動作クラスＢとして，提案手法によりプロトコル学習させた後，手話画像データベース全体の検索処理を行った．

検索結果における上位１０番までの候補とその評価値を表 4.4 に示す．また，各候補に対応する軌跡画像を図 4.9 に示す．

表 4.4 別クラスでの学習による類似手話動作の検索結果

「バイバイ」			「ダメ」		
No	動作名	評価値	No	動作名	評価値
1	バイバイ	36.00	6	上手	4.69
2	ダメ	34.67	7	学校	4.64
3	操る	5.75	8	注目	4.53
4	成長	5.18	9	頭にくる	4.49
5	平気	4.94	10	任す	4.24

第１位　第２位　第３位　第４位　第５位

第６位　第７位　第８位　第９位　第１０位

図 4.9 別クラスでの学習による類似手話動作の検索結果

図 4.9 において，第１位と第２位には学習対象の手話動作が挙げられているが，第３位以降は，右腕動作の候補や左腕動作の候補，さらに両腕動作の候補がバラバラに出現している．これは，別々の動作クラスでプロトコル学習させた場合，手話動作の視覚的な共通項に基づく検索処理が行われないことが原因であると考えられる．なお，本実験での検索所要時間は 20.69[s]，平均処理速度は 41.91[fps] であった．

4.2.3 同一クラスでの学習による検索実験

本節では複数の手話動作を同一クラスでプロトコル学習させた後，検索処理を行い得られた結果とその考察を示す．特徴量に基づく学習に用いる標準画像列には「バイバイ」の画像サンプルを使用し,「バイバイ」と「ダメ」を共にプロトコル学習させた後，手話画像データベース全体の検索処理を行う．

検索結果における上位１０番までの候補と，その評価値を表 4.5 に示し，各候補に対応する軌跡画像を図 4.10 に示す． なお，本実験での検索所要時間は 15.35[s]，平均処理速度は 56.49[fps] であった．

「バイバイ」は右腕を振る動作,「ダメ」は左腕を振る動作として手話画像データベースに登録されているが，これらの動作を同一クラスとしてプロトコル学習させた結果，表 4.5 に示すように，どちらか一方の腕を振る動作である「上手」

表 4.5 同一クラスでの学習による類似手話動作の検索結果

「バイバイ」＋「ダメ」					
No	動作名	評価値	No	動作名	評価値
1	バイバイ	58.8	6	遠い	29.9
2	ダメ	40.1	7	情報	29.7
3	上手	33.0	8	勉強	29.6
4	空	31.7	9	注目	29.2
5	台風	30.9	10	ＯＫ	29.1

第１位　第２位　第３位　第４位　第５位

第６位　第７位　第８位　第９位　第１０位

図 4.10 同一クラスでの学習による類似手話動作の検索結果

と「空」が第３位，第４位の順にランク付けされている．提案手法により,「片腕を振る」という視覚的な共通項が重視されたため，このような結果が得られたと考えられる．

また，同一クラスで複数手話動作のプロトコルを学習させる場合，１動作分の標準パターンを登録しておけば良いので，パターン照合コストが増大しないメリットがある．実際に，前節での実験における検索所要時間が 20.69[s]（平均処理速度は 41.91[fps]）であるのに対して，本実験の場合は 15.35[s]（平均処理速度は 56.49[fps]）であり，約 25 (%) の検索時間の短縮が実現されている．以上の結果は，同一クラスでのプロトコル学習により，類似動作の視覚的な共通項に基づく検索が可能となるのみならず，高速な検索処理が実現されることを示している．

4.2.4 有効注視点数を変化させた場合の検索実験

本節ではこれまでの評価実験で３２種類に固定していた有効注視点数を変化させた後，検索処理を行い，その際に得られた結果とその考察を示す．具体的には「バイバイ」を提案手法により学習させた後，有効注視点数を｛２４，１６，８，１｝の４通りに変更し，画像データベース全体の検索処理を行った．なお，有効注視点数＝３２の場合は，表 4.2 における「バイバイ」の結果と同一であるため，本節では省略する．

各有効注視点数における上位５番までの候補とその評価値を表 4.6 に，検索所要時間と平均処理速度については表 4.7 に示す． 有効注視点数＝２４の場合の検索結果の各候補に対応する軌跡画像を図 4.11 ，有効注視点数＝１６の場合の検索結果の各候補に対応する軌跡画像を図 4.12 ，有効注視点数＝８の場合の検索結果の各候補に対応する軌跡画像を図 4.13 ，有効注視点数＝１の場合の検索結果の各候補に対応する軌跡画像を図 4.14 にそれぞれ示す．

表 4.6 に示すように，有効注視点数が｛２４，１６，８｝においては，候補順位が変化しても同じ候補内容となっているが，有効注視点数＝１となると候補内容は大きく変化している．前者の場合は複数の注視点が考慮されている一方で，後者の場合は差分画像特徴の注視点しか考慮されていないことが，最大の原因であると考えられる．また，この結果は,「バイバイ」の検索では有効注視点数は８程度で十分であることを示している．正確な検索に必要となる注視点数は事前に

表 4.6 有効注視点数を変化させた場合の検索結果

有効注視点数＝２４			有効注視点数＝１６		
No	動作名	評価値	No	動作名	評価値
1	バイバイ	27.00	1	バイバイ	18.00
2	任す	3.98	2	任す	2.55
3	上手	3.16	3	ＯＫ	2.40
4	ＯＫ	3.02	4	上手	2.14
5	姉	2.74	5	姉	2.07

有効注視点数＝８			有効注視点数＝１		
No	動作名	評価値	No	動作名	評価値
1	バイバイ	9.00	1	バイバイ	1.125
2	任す	2.55	2	飲む	4.83E-2
3	ＯＫ	2.40	3	強い	4.04E-3
4	上手	2.14	4	集まる	4.04E-3
5	姉	2.07	5	つまり	0.59E-3

表 4.7 有効注視点数を変化させた場合の検索性能

有効注視点数	所要時間 [s]	平均処理速度 [fps]
32	15.16	57.16
24	12.94	67.00
16	9.49	91.37
8	6.06	143.00
1	4.23	205.02

予測できないことから，初期状態では有効注視点を多めに設定しておく必要があるが，表 4.6 からも分かるように，提案手法においては，冗長な注視点の設定による悪影響がほとんど見られない．本実験結果より，冗長な注視点設定下でも安定した検索結果を得られることが分かった．

なお，有効注視点数＝１の場合においても妥当な検索結果が得られており，その場合の平均処理速度が有効注視点数＝３２の場合の３．６倍に達し，１分間当り約１２，０００枚の手話画像（約７分相当のビデオレート動画像）を処理可能であるメリットがある．以上の結果は，検索の際に使用する有効注視点は多ければ多いほど良いということではないことを示す一方で，最終的な検索結果の妥当性については，利用者の判断に任せるのも有効な策の一つであることを示唆している．

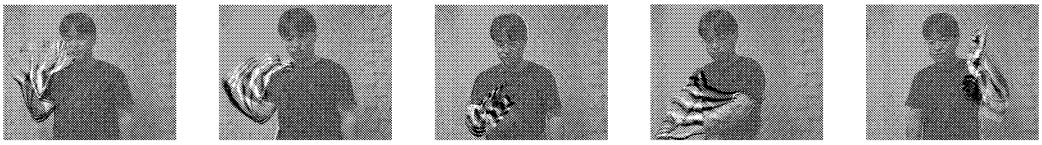
次に，任意の制御指標 S におけるシステム全体の処理速度と内部の処理速度を図 4.15 に示す．図 4.15 の結果からも分かるように，システム全体での処理速度は制御指標 S を低下させても大幅な向上は見られない．本評価実験により，多注視点選択制御におけるパターン走査間隔 S_k とパターン照合間隔 RS_k の値域を $[1, 3]$ 程度にあらかじめ制限しておくことは妥当であることが明らかになった．

第１位　第２位　第３位　第４位　第５位

図 4.11　有効注視点数＝２４の場合の検索結果

第１位　　第２位　　第３位　　第４位　　第５位

図 4.12　有効注視点数＝１６の場合の検索結果



第１位　　第２位　　第３位　　第４位　　第５位

図 4.13　有効注視点数＝８の場合の検索結果

第１位　　第２位　　第３位　　第４位　　第５位

図 4.14　有効注視点数＝１の場合の検索結果

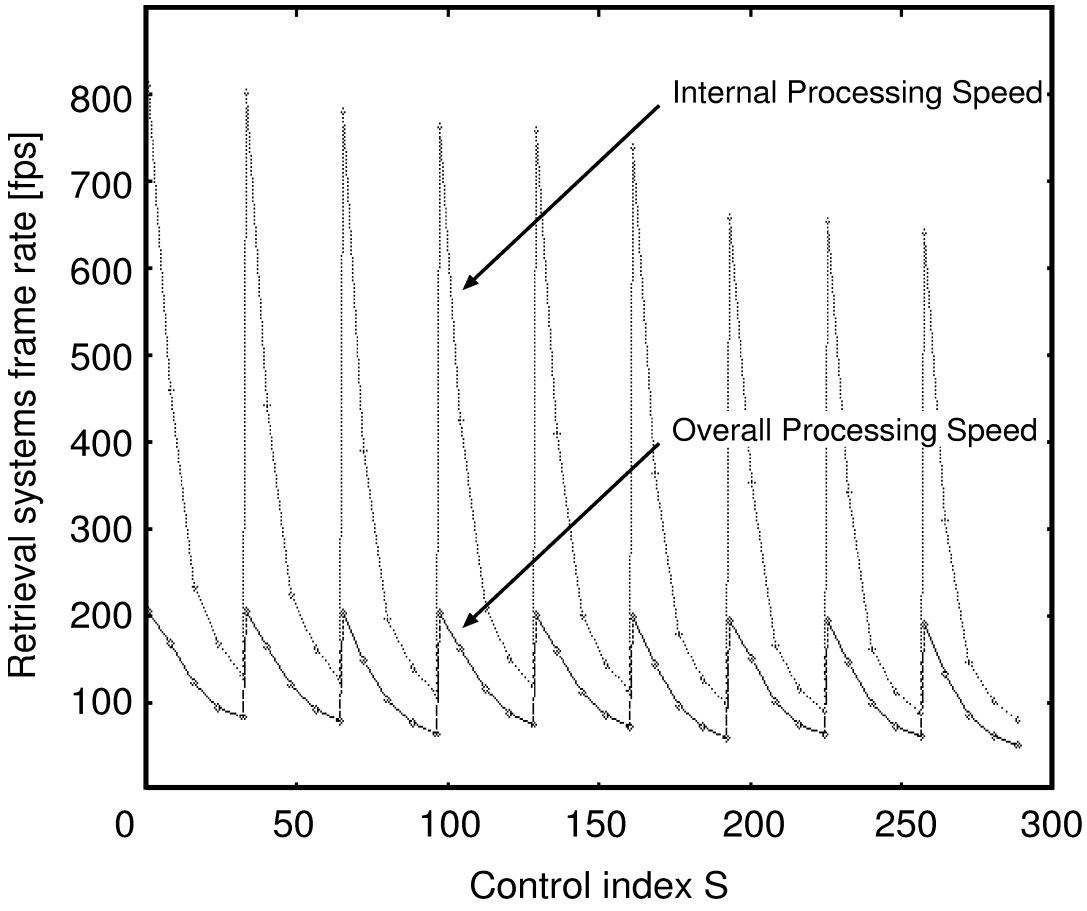


図 4.15 システムフレームレートの変化

4.2.5 任意指定時間での検索実験

本節では複数の手話動作を同一クラスとしてプロトコル学習させた後，３通りの指定時間での検索実験を行い，その際に得られた結果，および，その考察について述べる．本実験では，多注視点選択制御により，指定時間での類似手話動作検索を行えることを示す．具体的には，８秒・１２秒・１６秒の３通りの指定時間で，手話画像データベース内の全８６４枚の画像に対する検索処理を行う．実験は以下の手順に従って行う．

［１］複数の手話動作としては「バイバイ」と「ダメ」を採用し,「バイバイ」の画像サンプルを使用した特徴量に基づく学習の後，これらの動作を同一クラスとしてプロトコル学習させる．

［２］指定された時間から，理想的な処理速度を求め，検索システムにフレームレートを設定する．

［３］多注視点選択制御を開始する．

［４］手話画像データベース検索を開始する．

［５］多注視点身振り認識法により画像列を評価し，候補動作名と評価値，および，各画像列の処理に要した時間を記録する．検索が終了するまで手順［５］を繰り返す．

［６］検索終了時の各操作量 RS_k，S_k，N_{vip}，および，認識処理の内部構成グラフを記録する．

図 4.16 に，プロトコル学習の結果得られたプロトコルマップとその注視点重みを示す．図 4.16 において，左側の列には，各注視点の見出し（Diff. →差分画像を利用する注視点，Silh. →シルエット画像を利用する注視点，Edge. →エッジ画像を利用する注視点，P →位置，S →スケール，R →回転，0 →非依存，1 →依存）を示し，中央列にプロトコルマップ，さらに右側の列に注視点重みを示した．

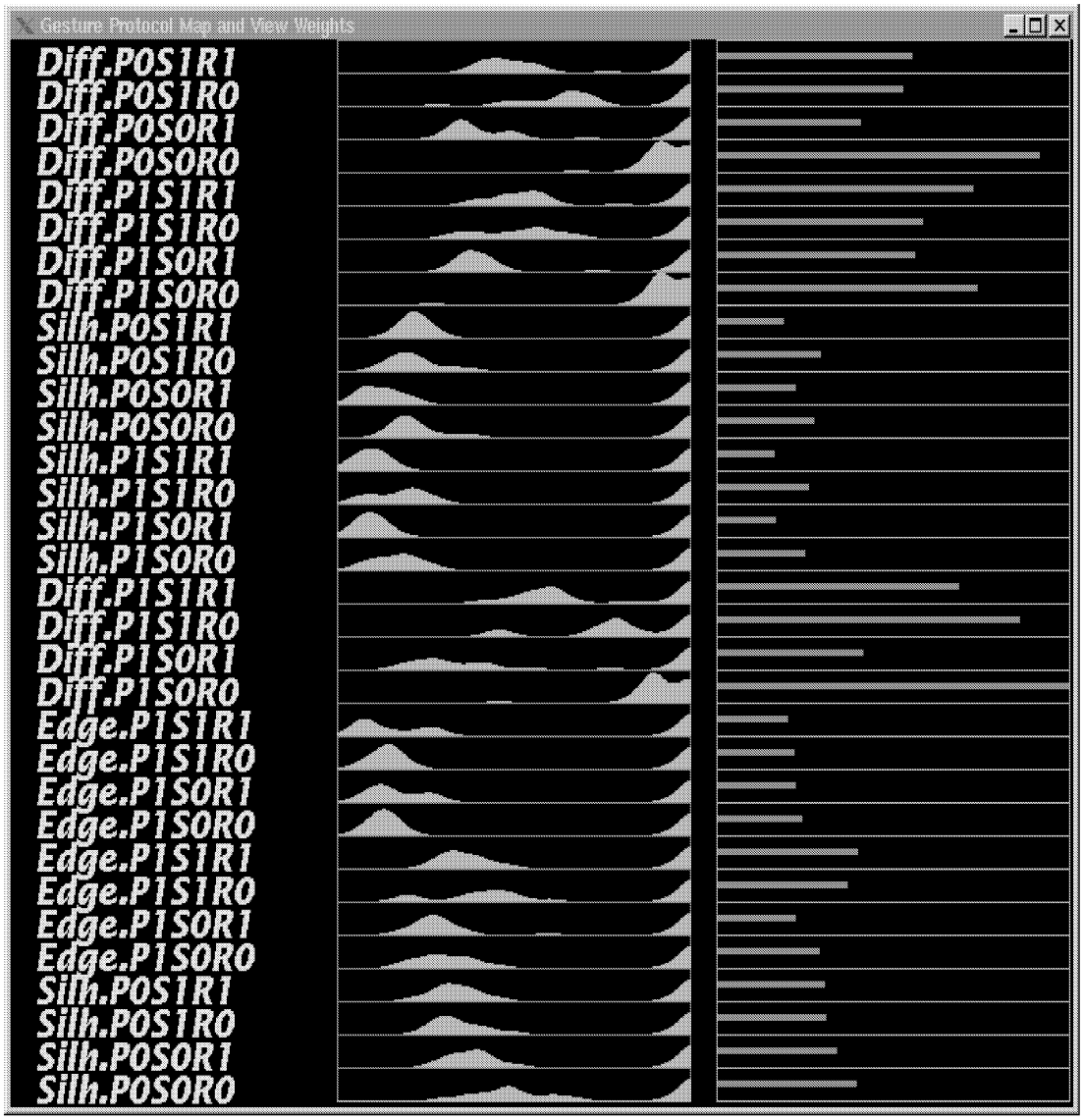


図 4.16　プロトコルマップと注視点重みの生成結果

本実験では，目標フレームレートの設定により指定時間での検索処理を行うが，具体的な目標フレームレートは式 (4.1) により求めている．

$$\text{Target Frame Rate[fps]} = \frac{\text{Number of Search Images[frames]}}{\text{Available Search Time[s]}} \tag{4.1}$$

手話画像データベースに登録されている全画像枚数は８６４枚であるが，特に動作の開始時と終了時には差分量が検出されず無効となる画像が存在している．無効となる画像の枚数は，差分画像を２値化する際のしきい値の影響を受ける．このしきい値を高くすればするほど手話動作の細部が無視される一方で，低くすればするほど照明の変動や衣服の微小や動きによるノイズの影響を受けるため，差分画像を２値化する際のしきい値は経験的に設定している．このしきい値を３０に設定した結果，本実験における有効画像枚数は７１３枚となることを事前に確認した．なお，無効となる画像についても有効となる画像と同様に特徴抽出処理などが実行されるため，処理時間が０になることはない．

式 (4.1) により，８秒間での検索の場合，目標フレームレートは $\frac{713}{8} \approx 89$[fps] に設定すれば良いことが分かる．同様にして，１２秒間の場合は $\frac{713}{12} \approx 60$[fps] となり，１６秒間の場合は $\frac{713}{16} \approx 45$[fps] に設定すれば良い．なお，目標フレームレートの設定以降は，自動的に認識系の再構成が多注視点選択制御法により行われる．

上記手順に従って得られた上位１０番までの候補動作名とその評価値および導出された認識系の内部構成グラフを，時間制御なしの場合については表 4.9 と図 4.21 に，１６秒検索については表 4.10 と図 4.22 に，１２秒検索については表 4.11 と図 4.23 に，８秒検索については表 4.12 と図 4.24 にそれぞれ示す．また，時間制御なしの場合の検索結果の各候補に対応する軌跡画像を図 4.17 ，指定時間（１６秒）における検索結果の各候補に対応する軌跡画像を図 4.18 ，指定時間（１２秒）における検索結果の各候補に対応する軌跡画像を図 4.19 ，指定時間（８秒）における検索結果の各候補に対応する軌跡画像を図 4.20 にそれぞれ示す．

表 4.9 ～表 4.12 に示した結果より，検索結果の候補動作については，時間制御なしの場合，１６秒検索の場合，１２秒検索の場合での結果が同様であるのに対し，８秒検索での結果は，他と比べて著しく異なっていることが分かる．８秒

検索において考慮される注視点は５種類程度であるのに対して，その他の条件では１０種類以上の注視点が考慮されている．冗長な注視点が設定されている場合，検索結果は類似した内容となる一方で，必要不可欠な注視点が無効とされている場合，検索結果は異なった内容になることが，本実験により明らかになった．

必要不可欠な注視点は，検索対象の手話動作に依存し，あらかじめ予測することが困難である上に，無用な注視点と必要不可欠な注視点を明確に区別することも実際には困難である．従って，これらの検索結果は，常に利用者に分かりやすい形態でフィードバックされなければならない．条件や設定を変更した上で，さらに検索を継続するかどうかは，利用者の判断に委ねるのが現状では最良の選択であると考える．手話画像データベースの検索では，利用者さえもシステムの一構成要素として位置付ける視点が不可欠である．

表 4.8 に各指定時間での検索実験における所要時間の計測結果とその誤差を示す．１２秒検索と１６秒検索の場合，ほぼ目標どおりの時間で検索処理を終えているが，８秒検索の場合には，7(%) 程度の誤差が生じている．本実験では，検索処理の最中においても認識系の再構成を含めた選択制御が行われるため，最終的な検索時間には選択制御に要した時間も含まれている．８秒検索の場合，表 4.8 に示すように，制御指標 S が２８８から１０１まで低下する必要があったため，他の場合よりも制御時間を要し，上述の結果になったと考えられる．この結果は，多注視点選択制御の速応性を向上させることが，検索時間の短縮と誤差の解消につながることを示唆している．

表 4.8 指定時間検索における所要時間とその誤差

目標検索時間 [s]	検索時間 [s]	誤差 (%)	制御指標 S
制限なし	14.33	—	288
16	15.93	0.4	284
12	11.91	0.8	269
8	8.58	7.3	101

なお，図 4.22 〜図 4.24 に示した認識系の内部構成グラフは，検索処理中の負荷状況に応じて導出されたものであり，

［１］検索対象の手話動作を増やす場合，

［２］本実験と異なる計算機環境において実験を行う場合，

［３］ＯＳ環境下において，提案システム以外にも大きな負荷が存在する場合

などでは本実験とは異なる結果が導かれる．提案手法は，検索システムの置かれる状況に柔軟かつ動的に対応し，場合によっては認識系を再構成することにより，最適な検索性能が得られるよう動作している．

以上の結果は，多注視点選択制御により指定時間でのデータベース検索が可能となることを示している．こうした機能は，インターネット上での画像検索エージェントを実現する際にも有効であると考えられる．

表 4.9 時間制御なしの場合の検索結果

No	動作名	評価値	No	動作名	評価値
1	バイバイ	58.78	6	遠い	29.94
2	ダメ	40.06	7	情報	29.74
3	上手	33.04	8	勉強	29.66
4	空	31.65	9	注目	29.19
5	台風	30.90	10	ＯＫ	29.18

表 4.10　１６秒検索の結果

No	動作名	評価値	No	動作名	評価値
1	バイバイ	58.77	6	遠い	29.94
2	ダメ	40.06	7	情報	29.74
3	上手	33.04	8	勉強	29.66
4	空	31.65	9	ＯＫ	29.18
5	台風	30.90	10	やっと	28.74

表 4.11　１２秒検索の結果

No	動作名	評価値	No	動作名	評価値
1	バイバイ	15.79	6	思う	8.03
2	ダメ	10.58	7	ＯＫ	7.64
3	台風	9.48	8	空	7.63
4	上手	8.20	9	姉	7.56
5	任す	8.15	10	遠い	7.45

表 4.12　８秒検索の結果

No	動作名	評価値	No	動作名	評価値
1	音楽	44.53	6	スポーツ	26.79
2	台風	40.51	7	生まれる	23.11
3	平等	37.41	8	空	22.37
4	木曜日	33.34	9	体育	18.13
5	値上げ	30.54	10	情報	17.74

第１位　第２位　第３位　第４位　第５位

第６位　第７位　第８位　第９位　第１０位

図 4.17　時間制御なしの場合の検索結果

第１位　第２位　第３位　第４位　第５位

第６位　第７位　第８位　第９位　第１０位

図 4.18　１６秒検索の結果

第１位　第２位　第３位　第４位　第５位

第６位　第７位　第８位　第９位　第１０位

図 4.19　１２秒検索の結果

第１位　第２位　第３位　第４位　第５位

第６位　第７位　第８位　第９位　第１０位

図 4.20　８秒検索の結果

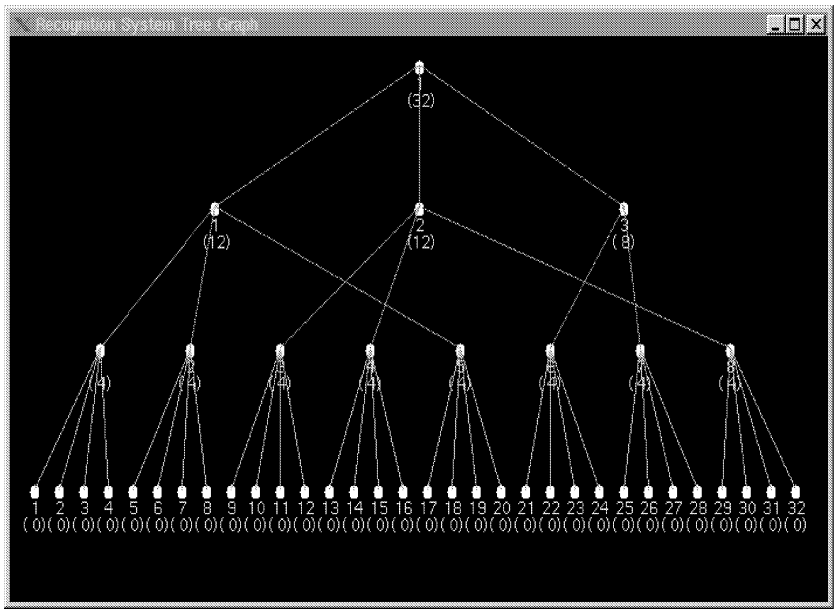


図 4.21　時間制御なしの場合の認識系の内部構成グラフ

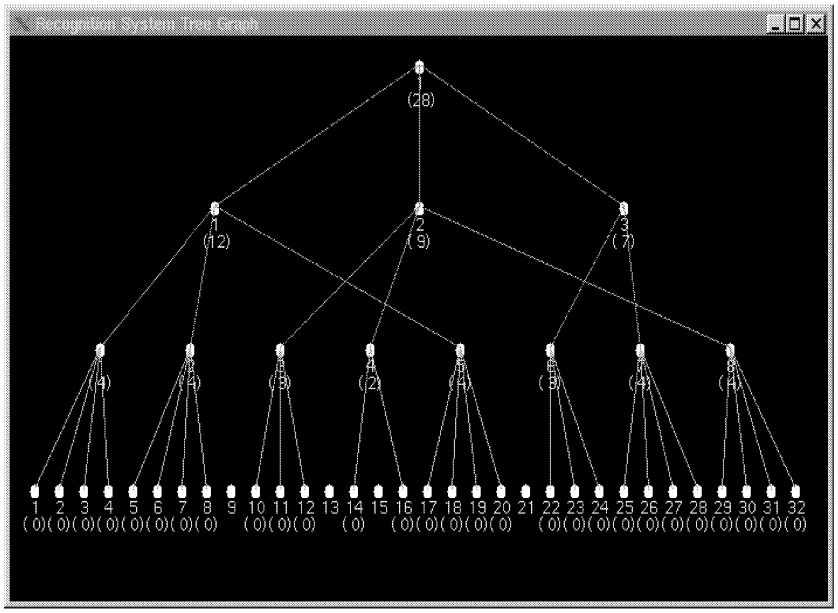


図 4.22　１６秒検索での認識系の内部構成グラフ

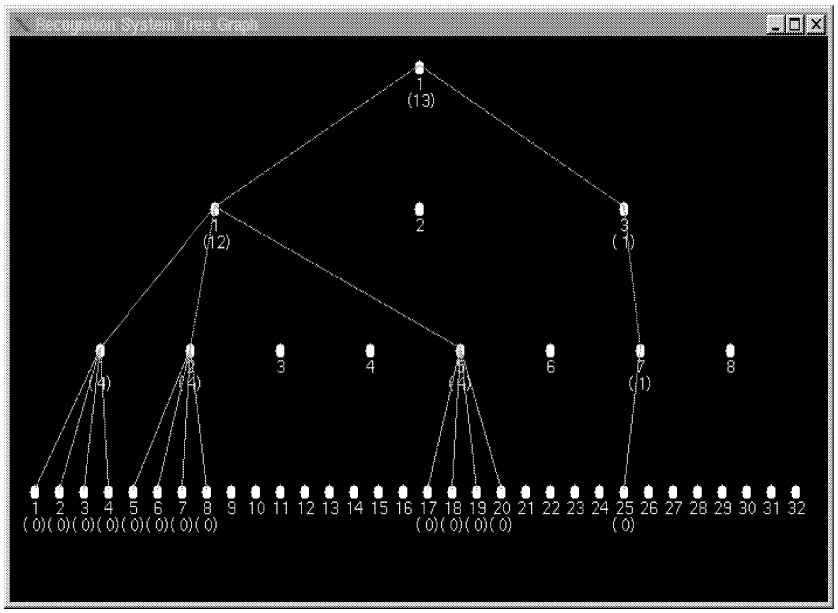


図 4.23　１２秒検索での認識系の内部構成グラフ

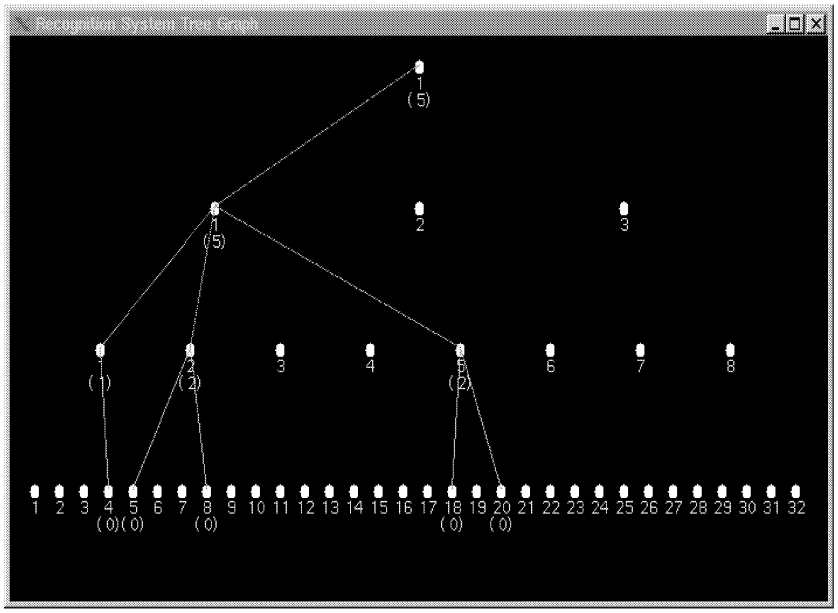


図 4.24　８秒検索での認識系の内部構成グラフ

5 考察・検討

5.1 多注視点身振り認識法について

5.1.1 身振りプロトコルの学習

評価実験では，提案手法により身振りプロトコルの学習とそれに基づく認識が可能になることを示した．具体的に身振り「バイバイ」と「グッパ」について，シルエット画像特徴と差分画像特徴を同時に考慮する方が，どちらか一方のみを考慮するよりも高い認識率が得られることを示した．この結果は，複数種類の特徴量を同時に考慮する提案手法の有効性を端的に示している．

プロトコル学習では学習対象の動作に依存して注視点重みが設定される．このため，システム開発者による最適な特徴量の選択が不要となり，認識システムを開発する際のコストを大幅に削減することが可能である．一方，プロトコル学習の収束性評価実験において，数個の訓練サンプルで十分な学習効果が得られることを示した．類似動作に対する学習曲線が類似していることから，プロトコル学習の再現性は十分に確保されていることが分かる．

以上の結果は，提案手法により利用者の意図を反映した学習と，不必要な情報に対して頑健な認識処理が実現されることを示している．もし，認識処理の階層化によりシステムの柔軟性が高まり，その結果として新たな身振り認識問題への対応が可能となるのであれば，さらなる階層化の可能性も検討に値するであろう．しかし，認識手法をむやみやたらと階層化するのでは，それぞれの階層の位置付けや意義・役割が見失われてしまうことにもなりかねず，階層化に際してはどのような新しい問題に対応可能となるのかを事前に明らかにしておく必要がある．

5.1.2 衣服依存性と身振り依存性評価

任意被験者の類似身振りサンプルをプロトコル学習させることにより，６種類の衣服での身振りに対する認識率が平均 30(%) 程度改善されることを示した．６種類の衣服すべてについて認識率の改善が認められたことから，プロトコル学習により衣服に比較的依存しない認識処理が可能となることが分かった．一方，身振り依存性についても，認識の困難であった黒のジャージによる身振りが，プロトコル学習により認識可能となることを示した．この結果は，プロトコル学習により認識処理の衣服依存性のみならず身振り依存性についても改善することが可能であることを示している．

以上の結果は，利用者の着衣と身振りの性質を限定しない身振りインタフェースが提案手法により実現可能であることを示している．なお，プロトコル学習では身振り標準パターンの増加がないため，認識処理速度低下の問題が生じないことも大きなメリットの一つである．

5.1.3 個人依存性評価

６種類の衣服を同一被験者に着用させて作成した訓練サンプルをプロトコル学習させた後，１１名の被験者の未知テストサンプルに対して認識実験を行った．従来研究では，被験者の動作位置や姿勢のみならず，服装や性別，さらに動作速度などに何らかの事前制約を設けるのが一般的であるが，本実験ではそうした制約を一切設けなかったにも関わらず，約 70(%) の平均認識率を得た．この結果は，複数種類の衣服を同一被験者に着用させて作成した訓練サンプルをプロトコル学習させることにより，他者身振りの認識率が改善されることを示している．

他の被験者の訓練サンプルを直接プロトコル学習させれば本実験を上回る認識率が得られると考えられるが，同一被験者の身振りが他の被験者の身振りと何らかの共通点を持つことを明らかにするために，本実験では他者身振りの訓練サンプルは使用しなかった．本実験の特徴を以下に列挙する．

［特徴１］ 被験者にキーワードで身振り動作を求めている．

［特徴２］ 被験者の位置・姿勢・動作に関して事前制約を設けていない．

［**特徴３**］被験者の性別や体格に事前制約を設けていない．

［**特徴４**］被験者の服装に事前制約を設けていない．

［**特徴５**］身振り動作の程度や速度について事前制約を設けていない．

本実験では，身振りの種類や軌跡の大小に依存しない認識処理のみならず，身振り動作の個人間格差や変動に対して頑健な認識処理が提案手法により実現されることを示した．認識の対象となる身振りに事前制約を設けない提案手法は，画像認識による身振りインタフェースに不可欠である．

5.1.4 身振り情報の推定と認識処理の実時間性評価

提案手法により身振り位相値を実時間かつフレーム毎に推定できることを示した．身振り位相値により，動作の向きや身振り速度，さらに身振り振幅を同時に推定することが可能となった．提案手法は，利用者の動作に即座に応答する仮想現実感システムを構築する際に有効である．

しかしながら，提案手法における認識処理速度が身振り標準パターンの増加に反比例する問題がある．このことは，仮想現実感応用などにおける高速な認識処理への要求を満たそうとすれば，認識の対象とする身振りの数に制限を設ける必要が生じることを意味している．平均的な身振りは１～２秒程度の動作を伴う．もし，ビデオレートでこうした動作を学習・認識させるのであれば，１動作につき３０～６０枚程度の標準画像列を扱うことが要求されることになる．図 2.24 より処理速度がビデオレート付近になるのは，１６種類の注視点で標準画像列が５０枚程度の時であるので，この場合１～２種類程度の動作クラスしか認識対象にできないことになる．

仮想物体操作には，少なくとも我々が短時間で記憶できる６種類程度の動作クラスの認識が必要になると仮定すれば，提案手法ではこの要求さえも満たせないことになる．本実験の結果は，提案手法における上記問題点を提起し，認識処理の負荷を直接制御する多注視点選択制御法を考案する発端となった．

なお，上述の標準パターンの増加による認識処理速度低下の問題は，図 3.18 に示したように多注視点選択制御法により克服した．

5.2 多注視点選択制御法について

5.2.1 応用システムとの接続

応用システムとしてジェスチャービデオシステムを開発した．ジェスチャービデオシステムは，利用者の瞬時瞬時の動作に対応する画像を即座に表示する仮想現実感アプリケーションの一例である．提案手法により実装した身振りインタフェースシステムを用いて利用者の身振り情報を認識させ，それらをジェスチャービデオシステムに伝送することにより，実時間での任意ビデオ画像とのインタラクションを実現した．

本実験では，身振りインタフェースシステムとジェスチャービデオシステムのプロセス間通信接続により生じる認識処理速度の低下および不安定化の問題の克服には提案手法が有効であることを示した．具体的には，17[fps] 〜 21[fps] で変動していた処理速度を，提案手法により常時 30[fps] へと安定化させることが可能であることを示した．この結果は，ビデオレートでの仮想体験が提案手法により可能になったことを示している．本実験では 17[fps] から 30[fps] へと安定化させるまでに約４秒の時間を要した．今後，制御に要する時間を極力短縮することにより，提案手法の有効性をさらに高める必要がある．

なお，身振りインタフェースシステムにより提供される身振り情報は，広範かつ多岐に利用されると考え，アプリケーションシステム開発の便宜を図るためのライブラリ関数群の整備を行った．

5.2.2 認識率への影響評価

提案手法により身振り認識手法のフレームレート依存性評価が可能となったため，１０通りのフレームレート条件（1 〜 45[fps]（5[fps] 間隔））の下での手話動作の認識率変化を調べた．認識率に影響を及ぼす第一要因は，認識対象動作の数ではなく，むしろ，その動作の性質にあると考え，動作軌跡の異なる３つのグループ，各８種類の手話動作を対象にした．

実験の結果，３つのグループ全体で約 96(%) の平均認識率を得た．このことは，今回選定した動作群については提案手法により認識率を高く維持した上で任意の

処理速度が実現されていることを示している．また，各手話動作の平均認識率についても，一部の例外を除いて 90(%) 以上の平均認識率が得られた．

なお，35[fps] を超える領域では，平均認識率の有意な低下を確認したが，これは NTSC 規格に準拠した画像入力装置を使用したことが原因であると考えられる．また，35[fps] を超える領域では制御指標 S の急激な低下を確認したが，これは画像入力に必要な時間が一定であるために認識処理側でより一層の負荷削減が要求されたことを示しており，ここに至って NTSC 規格に準拠した画像入力装置がボトルネック要因として無視できなくなることが分かった．

5.2.3 認識処理の実時間性評価

身振り標準パターンの増加に伴う照合コスト増大の問題には提案手法が有効であることを評価実験により示した．制御ありの場合，身振り標準パターンが増加しても常に 30[fps] での認識処理を実現する一方で，制御なしの場合，処理速度は身振り標準パターンの増加に反比例して低下して行く．本実験では，認識対象の身振りを増やす場合でも常時 30[fps] の処理速度が実現されることを示した．この結果は，提案手法の有効性を最も端的に示している．なお，提案手法は任意速度での認識処理を実現するため，アプリケーションシステムからの様々な動作速度への要求にも応えることができる．

一方，提案手法は認識システムの処理速度を逐次把握しながら負荷制御するため，動作速度の異なる計算機環境においても，任意速度での認識処理を実現できるメリットがある．高速な計算機の場合，同一処理速度条件の下にあっても，低速な計算機より多くの注視点が考慮される違いがある．計算機性能の違いは提案手法により把握され最適な有効注視点数が決定されるため，利用者のみならず開発者も計算機環境の違いを意識する必要がない．

以上の結果は，提案手法により，仮想現実感のための身振りインタフェースをより実用的なレベルにおいて構築できることを示している．

5.3 手話画像データベース検索システムについて

5.3.1 類似手話動作の検索

任意の手話画像サンプルをプロトコル学習させることにより，手話画像データベースから類似手話画像サンプルの動作名を検索できることを示した．具体的には，４種類の手話画像サンプルを用いて検索実験を行った．その結果，片腕動作のプロトコル学習をした場合には，片腕動作の手話画像サンプルの動作名が得られ，両腕動作のプロトコル学習をした場合には，両腕動作の手話画像サンプルの動作名が得られることが分かった．

提案システムでは，動作の奥行き・顔表情・視線方向・手形状など手話動作の認識にとって重要な要素を考慮していないものの，大雑把な身体動作から類似動作サンプルの動作名を検索する際には有効である．もし，動作の奥行き・顔表情・視線方向・手形状などを考慮する必要がある場合には，各要素に対応する認識モジュールを提案手法に組み込むことで対応できる．これら認識モジュールの選択作業は，提案手法により全自動かつ実時間で実行される．以上のことは，提案手法が多様な認識モジュールを統合するための一般的枠組みとして活用できることを意味している．

なお，選択制御なしの場合でもビデオレートを上回る平均処理速度が得られており，高速な手話画像データベース検索が提案手法により実現された．

5.3.2 別クラスおよび同一クラスでの学習による検索

手話画像データベース検索におけるプロトコル学習の役割とその意義を明らかにするため，別クラスでの学習の場合と同一クラスでの学習の場合での検索実験を行った．本実験では，片腕動作「バイバイ」と「ダメ」に類似した手話動作を検索させた．

別クラスでの評価実験では，２種類の手話画像サンプルを別々の動作クラスに登録し学習させた．この結果，期待はずれの両腕動作の候補が上位にランク付けされた．別クラスでの学習では１種類の手話画像サンプルしかプロトコル学習させなかったため,「片腕の動作が重要」というプロトコル情報が獲得されなかった

ことがこの原因として挙げられる．

同一クラスでの評価実験では，２種類の手話画像サンプルを同一の動作クラスとして登録し学習させた．この場合，期待通りに右腕または左腕動作の候補が上位にランク付けされることを確認した．同一クラスでの学習では，２種類の手話画像サンプルをプロトコル学習させたため，片腕の動作を重視した検索処理が実行されたと考えられる．

以上の結果は，複数の手話画像サンプルをプロトコル学習させることにより，部分的な特徴を選択的に重視した検索処理が実現されることを示している．また，プロトコル学習により，右腕と左腕を区別させたり区別させなかったりと，より概念的な検索処理が実現され，利用者の意図を検索処理に反映させることが可能となる．さらに，同一クラスでプロトコル学習させる場合，１動作分の身振り標準パターンの登録で済むため，平均処理速度が別クラスでの学習の場合よりも高速になるメリットがある．訓練サンプルから視覚的共通項を見出し，それらを重視した認識処理を実現する提案手法は，手話画像データベース検索課題においても有効である．

5.3.3 有効注視点数を変化させた場合の検索

有効注視点を｛２４，１６，８，１｝の４通りに変化させた場合，検索結果に変化が現れるかどうかを調べた．その結果，有効注視点数が２４から８までは候補内容がほとんど変化しないが，有効注視点数が１になると候補内容に変化が現れることが分かった．この結果は，本実験で対象とした「バイバイ」については８種類程度の有効注視点が必要となることを示す一方で，冗長な注視点の設定下でも提案手法により安定した検索結果が得られることを示している．

設定した注視点が冗長であるかどうかは対象動作の性質に依存するため，最適な注視点数を事前に予測することは困難である．１種類の注視点のみですべての対象動作を認識できる場合があれば，３２種類の注視点を駆使しても認識できない場合もあると考えられ，３２種類の注視点で何種類の身振りを認識できるかは，実際に試してみないと分からないのが現状である．ただし，故意に冗長な注視点を設定した上で，学習結果に基づき最適な注視点数を推定する手法も考えられ，この問題についてはさらなる検討の余地がある．

5.3.4 任意指定時間での検索

任意指定時間での類似手話動作検索が提案手法により可能となることを評価実験により示した．指定された時間内に利用可能な計算機資源を最大限に活用した検索処理を行うことから，この検索方式を Best-Effort 型検索と呼びたい．評価実験の結果からも明らかであるように，提案手法は手話画像の検索を高速かつ柔軟に行うことが可能であり，手話画像データベースの検索に有効である．

しかしながら，現状の Best-Effort 型検索では時間的な制約を満たすことを最優先するため，本来不可欠な注視点をも無効にしてしまい，８秒検索の結果に見られるような妥当性を欠いた検索結果を出力する場合がある．処理速度を追求すればするほど柔軟性が犠牲になり，柔軟性を追求すればするほど処理速度が犠牲となってしまうというジレンマが存在している．これは利用可能な計算機資源が有限である認識システムにおいて不可避の問題であり，検索処理の柔軟性と高速性の完全な両立が困難であることを示唆している．必要最小限の注視点数を事前に予測できれば上述の問題点を回避することは可能であり，今後この課題に取り組むことで最適な処理速度と認識性能の下での検索処理を実現する必要がある．

多くの課題が山積しているが，提案手法は認識システムに自己の負荷を自律的に制御させ，最適な動作状態に自力で到達させる試みの第一歩であることは確かである．提案システムを今後一層広範囲の手話画像検索ニーズに対応させるためには，図 5.1 に示すような検索モデルに基づいて，さらに研究を進めて行く必要がある．図 5.1 に示した検索モデルでは，検索処理を，

［レベル１］認識評価値に基づく手話動作の検索処理

［レベル２］位相値と速度の関係，位相値と振幅の関係，位相値と手話クラスの関係など，身振り情報を考慮した位相値空間における手話動作の検索処理

［レベル３］位相値空間での検索処理の結果を考慮した，状態空間における手話動作の検索処理

の３階層に分離・独立させる．こうすることにより，例えば「手話動作の向きと速度を考慮した検索」のように，より細かく条件を指定した上での検索が可能になる．

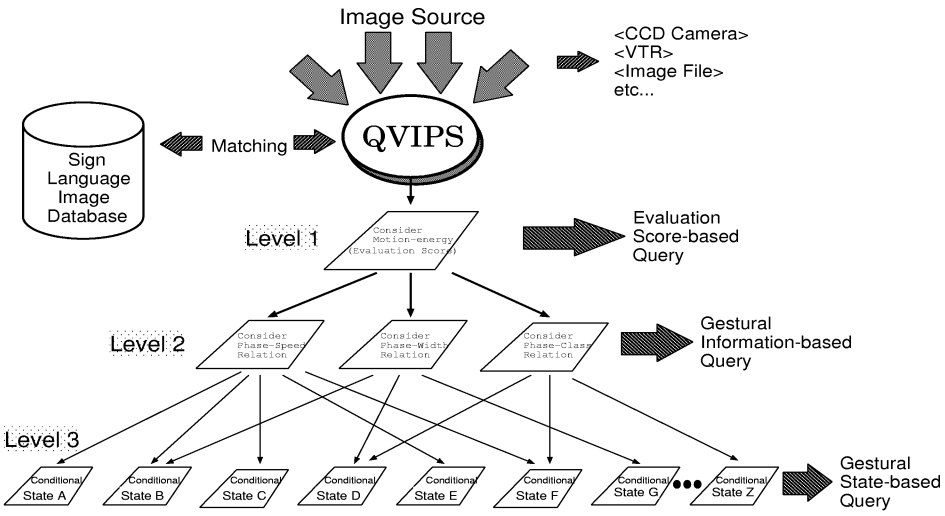


図 5.1 手話画像データベース検索モデル

5.4 身振りインタフェースの実用化に向けて

現在，一般に入手可能な画像入力装置の多くは，人間の視覚特性を根拠に標準化された NTSC などの規格に準拠している．このことは，コンピュータ視覚の研究をする際にほとんどの場合「人間用」視覚センシング装置の使用を余儀なくされていることを意味している．

ビデオレートは高度な補完能力を持つ人間の視覚系がよしとする速度であり，コンピュータに実世界の動きを認識させる際には，決して十分であるとは言えない．身振り動作には日常的な緩やかでゆっくりとした動作もあれば，スポーツ時の動作の様に非常に変化が激しく高速な動作もある．これら動作全般を認識の対象とする場合，さらに高速な画像入力装置を使用する必要がある．また，計測範囲を広げて認識処理を行う場合でも，現在の NTSC 規格では解像度が低過ぎる．NTSC 規格は既に広く普及しているという大きなメリットを有する一方で，コンピュータ視覚用のセンサとして利用する場合，上述のような問題点を抱えている．

一方，ソフトウェア面では，認識モジュールや階層数が増えると必然的に計算量が増大する問題がある．この問題に対処するには，並列化が可能な部分をマル

チプロセッサへと分散して処理する方式 [51] が有効であるが，各認識モジュールから得られる情報の統合方法，同期の取り方など，数多くの解決されるべき課題がある．

また，認識率などの主観的な指標は物理量として検出し定量化することが困難であり，認識システムへと実時間で動的にフィードバックさせることができなかった．今後こうした指標の定量化を自動的に行う手法の検討を進め，処理速度のみならず認識率についてもその品質が保証されるような身振りインタフェースシステムの実現を目指す必要がある．

6 結言

本論文では，非接触・非装着型身振りインタフェースの実現に不可欠である任意人物の任意身振りを画像により実時間で認識する手法を提案した．

まず，身振り画像の特徴選択問題を解決するために，多注視点身振り認識法を提案した．多注視点身振り認識法では，同一種類の身振りとして与えられる画像列から時空間的に安定した視覚的特徴群を見出し，それらを重視した認識処理を行った．評価実験では，多注視点身振り認識法による入力画像の位相値推定により，簡便な処理のみで身振り速度や身振り振幅などの身振り情報を求めることが可能であることを示した．また，多注視点身振り認識法は衣服や身振りの種類・軌跡の大小・動作特性の個人間変動に対して頑健な認識処理を実現することを明らかにした．多注視点身振り認識法により，身振りの種類や性質を事前に限定せず，広範な身振りを認識できるようになったことの意義は大きい．

続いて，身振り画像のサンプリング問題を解決するために，多注視点選択制御法を提案した．多注視点選択制御法によるパターン走査間隔とパターン照合間隔の選択制御ならびに多注視点の選択制御により，画像サンプリング間隔を任意の時間間隔に安定化させた．評価実験では，ビデオレートでの任意ビデオ画像とのインタラクションが多注視点選択制御法により実現されることを示した．また，処理負荷の増大や変動が発生する条件下でも，多注視点選択制御法により任意速度での身振り認識処理が実現されることを示した．多注視点選択制御法は，認識システムによる自律的な負荷制御機能の実現に向けた第一歩であり，その可能性と課題については今後さらに検討して行く必要がある．

上述の提案手法による画像検索機能の実現可能性を示すために，手話画像データベース検索システムを開発し，評価実験によりその有効性を実証した．提案システムでは，多注視点身振り認識法によるプロトコル学習に基づいて手話画像

データベースから類似手話動作に対応する動作名を検索させた．評価実験において，冗長な注視点の設定下でも安定した検索結果を得る一方，864 枚の検索対象画像に対して最短 4.2[s] での高速な検索処理が実現されることを示した．また，多注視点選択制御法により任意指定時間での画像データベース検索が可能になることを示した．従来困難であった同一人物の多様な動作の検索が，身振りにより直接実行できるようになったことの意義は大きい．

以上に総括したように，本論文では画像認識による身振りインタフェースの実現に不可欠な手法の提案のみならず，身振りインタフェースに画像検索機能を付与する試みの一例として，手話画像データベース検索システムを開発した．今後は，提案手法のさらなる改善と拡張を積極的に推進する一方で，画像検索機能を強化することにより，サイバースペースにおける身振りインタフェースの新たな役割とその可能性を追求して行きたい．

謝辞

本研究を遂行する機会のみならず長い期間に渡り常に適切なご指導と多大なるご助言を賜りました奈良先端科学技術大学院大学情報科学研究科教授千原國宏先生に心からの感謝の意を申し上げます．研究に不可欠な継続の意思，忍耐，集中力，機転など，研究者が堅持すべき姿勢の重要性を学ばせて頂きました．また，数多くの研究発表の機会を頂き，その中で数々の不足な点や問題点を指摘して頂いたことは，本研究を前進させる上で大いなる参考となりました．改めて厚く御礼申し上げます．

奈良先端科学技術大学院大学への入学当初以来，研究活動から学生生活に至るまで，的確かつ細やかな指導を賜るのみならず，研究の難しさから素晴らしさまでご指導頂きました大阪大学大学院基礎工学研究科助教授佐藤宏介先生に心から感謝の意を表します．研究に対する動機付けをより確なものとすることができました．重ねて御礼申し上げます．

本論文をまとめるに当たり，ご厚情に満ちたご指導と数々の不備のご指摘のみならず，常に適切なご教示を賜わりました情報科学研究科教授植村俊亮先生に厚く御礼申し上げます．

本論文の作成のみならず，シンポジウム・研究会等の機会を通じて，ご厚情に満ちたご指導，数々の有益なご教示を賜わりました情報科学研究科教授横矢直和先生に厚く御礼申し上げます．

本論文の作成にとどまらず，国際会議等の機会を通じて，ご厚情に満ちたご指導，数々の示唆に富むご教示を賜わりました情報科学研究科助教授竹村治雄先生に厚く御礼申し上げます．

日頃の研究活動のみならず，本論文の執筆にあたり，数々の的確かつ有益なご指導およびご助言を賜りました情報科学研究科助教授眞鍋佳嗣先生に厚く御礼申

し上げます.

本研究の過程において数々のご指導ならびにご助言を賜ったのみならず，具体的なご支援も頂きました先端科学技術研究調査センター助教授大城理先生に深く感謝の意を表します.

本研究を進めるに当り，日頃からの数多くのご助言，ご指導を頂きました東京大学新領域創成科学研究科助教授眞溪歩先生，ならびに和歌山大学工学部助教授陳謙先生に深く感謝致します.

学生生活から研究活動に渡り，必要不可欠な事項をご助言頂きました情報科学研究科助手土居元紀先生，情報科学研究科助手黒田知宏先生に心より感謝致します.

本研究を進めるに当たり，実験装置および機材の調達において，力強いご支援を頂きましたオムロン株式会社新事業開発センタファジィ推進室第１開発課主事川出雅人さん，松尾和幸さんに感謝致します.

奈良先端科学技術大学院大学における研究活動について深いご理解とご支援を頂きました奈良工業高等専門学校電気工学科教授高橋晴雄先生，同学科教授成田紘一先生，同校情報工学科助教授工藤英男先生，同学科助教授多喜正城先生に心から感謝申し上げます.

奈良先端科学技術大学院大学への入学以来，研究発表会や講義など，数々の場面でご指導ならびに有益なご議論・ご助言を賜るとともにお世話にもなりました情報科学研究科教官の先生方，ならびに事務・関係職員の皆様に心から感謝致します.

最後に，日頃から多方面に渡りご理解・ご協力ならびに力強いご支援を頂きました像情報処理学講座歴代秘書の皆様，像情報処理学講座の卒業生および現役生諸氏に心から感謝致します.

参考文献

[1] Alex P.Pentland : “Machine Understanding of Human Action”， 7th International Forum on Frontier of Telecommunication Technology，Tokyo，Japan，pp.110-119， November，1995

[2] Trevor Darrell，Alex P.Pentland : “Attention-driven Expression and Gesture Analysis in an Interactive Environment”， International Workshop on Automatic Face and Gesture Recognition，Zurich，pp.135-140，1995

[3] Roger E. Axtell : “GESTURES:The DO’s and TABOO’s of Body Language Around the World”， John Wiley & Sons Inc.，1991

[4] Paul Ekman : “Essential Behavioral Science of the Face and Gesture that Computer Scientists Need to Know”，International Workshop on Automatic Face and Gesture Recognition，Zurich，pp.7-11，1995

[5] 黒川隆夫：“ノンバーバルインタフェース”，オーム社，1994

[6] Ben Delaney，Michael J. Potel : “The Mystery of Motion Capture”，IEEE Computer Graphics and Applications，September/October，1998

[7] 大須賀節雄 : “知識工学講座 10 ヒューマンインタフェース”，オーム社，1992

[8] 横井茂樹：“ビジョンとグラフィクス”，培風館，1995

[9] Robert Aston，Joyce Schwarz : “Multimedia–Gateway to the Next Millennium–”， AP professional，pp.214-217，1994

[10] 山本正信，川田聡，近藤拓也，越川和忠：“ロボットモデルに基づく人間動作の 3 次元動画像追跡”，信学論（D-II），Vol.J79-D-II，No.1，pp.71-83，January，1996

[11] 佐藤明知，川田聡，大崎喜彦，山本正信：“多視点動画像からの人間動作の追跡と再構成”，信学論（D-II），Vol.J80-D-II，No.6，pp.1581-1589，June，1997

[12] 大田佳人，山際貴志，山本正信：“キーフレーム拘束を利用した単眼動画像からの人物動作の追跡”，信学論（D-II），Vol.J81-D-II，No.9，pp.2008-2018，September，1998

[13] 大垣健一，岩井儀雄，谷内田正彦：“動きと形状モデルによる人物の姿勢推定”，信学論（D-II），Vol.J82-D-II，No.10，pp.1739-1749，October，1999

[14] 米元聡，布巻崇，鶴田直之，谷口倫一郎：“多視点動画像における 3 次元多関節物体の追跡”，画像の認識理解シンポジウム（MIRU’98），講演論文集 Vol.2，pp.423-428，July，1998

[15] 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，July，1996

[16] Jun Ohya，Fumio Kishino：“Human Posture Estimation from Multiple Images using Genetic Algorithm”，7th International Forum on Frontier of Telecommunication Technology，Tokyo，Japan，pp.90-93，November，1995

[17] 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，August，1998

[18] 西村拓一，向井理朗，岡隆一：“複数人物によるジェスチャーの単一画像からのスポッティング認識”，信学技報，PRMU96-90，pp.77-84，November，1996

[19] 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，June，1997

[20] 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，January，1998

[21] 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，August，1994

[22] 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.1571-1581，August，1994

[23] 小俣寿之，奥野健太郎，秋田幸治，山口亨：“知的エージェントロボットにおける認識基礎学習”，12th Fuzzy System Symposium，pp.369-372，June，1996

[24] 大和淳司，大谷淳，石井健一郎：“隠れマルコフモデルを用いた動画像からの人物の行動認識”，信学論（D-II），Vol.J76-D-II，No.12，pp.2556-2563，December，1993

[25] 大和淳司，倉掛正治，伴野明，石井健一郎：“カテゴリー別ＶＱを用いたHMMによる動作認識法”，信学論（D-II），Vol.J77-D-II，No.7，pp.1311-1318，July，1994

[26] 畠直志，岩井儀雄，谷内田正彦：“動き情報と情報圧縮を用いたロバストなジェスチャ認識手法”，信学論（D-II），Vol.J81-D-II，No.9，pp.1983-1992，September，1998

[27] Thad Starner，Joshua Weaver，Alex Pentland：“Real-time American Sign Language Recognition using Desk and Wearable Computer Based Video”，IEEE Trans. Pattern Analysis and Machine Intelligence，Vol.20，NO.12，pp.1371-1375，December，1998

[28] 藤本泰史，岩佐英彦，横矢直和，竹村治雄：“固有空間内の軌跡の類似性に基づく動画像検索”，信学技報，PRMU96-110，pp.49-56，December，1996

[29] 大場光太郎，池内克史：“局所固有空間手法による金属物体の安定認識”，信学論（D-II），Vol.J80-D-II，No.12，pp.3147-3154，December，1997

[30] 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手法—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，June，1997

[31] Yuntao Cui，John Weng：“Learning-based Hand Sign Recognition”，Proceedings of the International Workshop on Automatic Face and Gesture Recognition，pp.201-206，Zurich，1995

[32] E.Hunter，J.Schlenzig，R.Jain：“Posture Estimation in Reduced-Model Gesture Input Systems”，Proceedings of the International Workshop on Automatic Face and Gesture Recognition，pp.290-295，Zurich，1995

[33] Simon X. Liao，Miroslaw Pawlak：“On the Accuracy of Zernike Moments for Image Analysis”，IEEE Trans. Pattern Analysis and Machine Intelligence，Vol.20，NO.12，pp.1358-1364，December，1998

[34] I.Scott MacKenzie，Colin Ware：“Lag as a Determinant of Human Performance in Interactive Systems”，Proceedings of the Human Factors in Computing Systems INTERCHI’93，pp.488-493，Amsterdam，April 24-29，1993

[35] Richard H.Y.So，Michael J.Griffin：“Experimental Studies of the Use of Phase Lead Filters to Compensate Lags in Head-Coupled Visual Displays”，IEEE Trans. Systems，Man，and Cybernetics–Part A:Systems and Humans，Vol.26，NO.4，pp.445-454，July，1996

[36] Thomas S.Huang，Vladimir I.Pavlovic：“Hand Gesture Modeling, Analysis, and Synthesis”，International Workshop on Automatic Face and Gesture Recognition，Zurich，pp.73-79，1995

[37] 井口征士，佐藤宏介：“三次元画像計測”，昭晃堂，1990

[38] 橋本浩一：“視覚フィードバック制御－静から動へ－”，システム/制御/情報，38(12)，pp.659-665，1994

[39] 一松信，村岡洋一：“感性と情報処理－情報科学の新しい可能性－”，共立出版，1993

[40] 安田稔: “視覚－人工化における諸側面”，数理科学 No.356，pp.10-16，February，1993

[41] 飯島泰蔵：“パターン認識”，電気・電子工学大系 43，コロナ社，pp.1-25，1973

[42] 松山隆司，久野義徳，井宮淳編：“コンピュータビジョン：技術評論と将来展望”，新技術コミュニケーションズ，pp.219-229，1998

[43] 國吉康夫：“実世界エージェントにおける注意と視点-情報の分節・統合・共有-”，人工知能学会誌 Vol.10 No.4，pp.507-514，July，1995

[44] 久野義徳：“アクティブビジョン－歴史と展望－”，人工知能学会誌 Vol.10 No.4，pp.493-499，July，1995

[45] 石黒浩：“注視に基づくロボットの視覚”，人工知能学会誌 Vol.10 No.4，pp.500-506，July，1995

[46] 横澤一彦: “特徴統合理論”，数理科学，No.351，pp.76-79，September，1992

[47] 喜多伸一: “視覚探索の神経機構”，数理科学，No.353，pp.64-67，November，1992

[48] 三浦利章：“外界情報の獲得・処理様式”，数理科学，No.354，pp.53-58，December，1992

[49] 伊藤政雄，竹村茂：“手話入門－ふれあいの言葉－”，廣済堂，1995

[50] 小早川倫広，星守: “画像内容に基づいた画像検索システム”，Ｂｉｔ，Vol.31，No.10，October，1999

[51] 小野定康，太田直久：“スーパシグナルプロセッシング”，デジタル信号処理シリーズ第 13 巻 昭光堂，1995

研究業績

(関連論文)

[学術論文誌]

[1] 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，May，1998

[2] 桐島俊之，佐藤宏介，千原國宏：“多注視点の選択制御による身振りの実時間画像認識”，信学論（D-II）（条件付採録：通知受領日 2000 年 3 月 6 日）

[国際会議]

[1] Toshiyuki Kirishima，Kosuke Sato，Kunihiro Chihara : “A Novel Approach on Gesture Recognition: The Gesture Protocol-based Gesture Interface”，Proceedings of International Conference on Virtual Systems and Multimedia(VSMM'96)，pp.433-438，September 18-20，1996

[2] Toshiyuki Kirishima，Kosuke Sato，Hirokazu Narita，Kunihiro Chihara : “Realtime Gesture Recognition under the Multi-layered Parallel Recognition Framework of QVIPS”，Proceedings of the third IEEE International Conference on Face and Gesture Recognition(FG'98)，pp.579-584，April，1998

［国内会議］

[1] 桐島俊之，佐藤宏介，千原國宏：”VR のための身振りのリアルタイム動画像認識”，画像の認識・理解シンポジウム（MIRU’96）講演論文集，Vol.II，pp.163-168，July 17-19，1996

[2] 桐島俊之，佐藤宏介，千原國宏：“多注視点の動的制御によるリアルタイム身振り認識”，画像の認識理解シンポジウム（MIRU’98），講演論文集 Vol.2，pp.19-25，July，1998

［研究会］

[1] 桐島俊之，佐藤宏介，千原國宏：“身振りの動画像認識によるユーザインタフェース”，信学技報，PRU95-190，pp.1-6，January 18-19，1996

（その他）

[1] 大久保修司，桐島俊之，佐藤宏介，千原國宏：“三次元動作認識における最適視点位置の選択”，平成１０年電気関係学会関西支部連合大会，G14-13，November，1998

[2] 桐島俊之：“視覚的条件付けによる身振りの実時間画像認識”，奈良工業高等専門学校研究紀要３３号，pp.29-34，1997

[3] 桐島俊之，佐藤宏介，千原國宏：“動的多注視点制御によるリアルタイム身振り認識”，奈良工業高等専門学校研究紀要３４号，pp.27-32，1998

[4] 桐島俊之，佐藤宏介，千原國宏：“手話画像データベース検索のための動画像認識”，奈良工業高等専門学校研究紀要３５号，1999

付録

A. 形状特徴抽出法

本付録では，式 (A..1) により表される形状特徴抽出オペレータのユニークさ（uniqueness）の証明，および，勾配係数 a の決定方法を示す．

$$q = \frac{R}{L} \frac{\sum_r p(r) \exp\left\{-a(r-\phi)^2\right\}}{\sum_r p(r)} \tag{A..1}$$

A..1 ガウス密度特徴の uniqueness について

式 (A..1) の目的は，線形パターン $p(r)$ からユニークな特徴量 q を算出することである．従って，畳み込み演算の際に使用するオペレータは，任意の位置 r においてユニークである必要がある．このような分布としては，

$$f(r) = ar \tag{A..2}$$

$$f(r) = \exp(-ar) \tag{A..3}$$

$$f(r) = \exp(-ar^2) \tag{A..4}$$

などが考えられる．使用するオペレータが対称性を持つ場合，左右対称なパターン分布からユニークな特徴量を抽出することが出来ないため，各オペレータの対称性を式 (A..5) により判定する必要がある．

$$\int_0^R \frac{d}{dr} \{f(r) + f(R-r)\}\, dr \begin{cases} = 0 & \text{(symmetry)} \\ \neq 0 & \text{(asymmetry)} \end{cases} \tag{A..5}$$

式 (A..2) のオペレータに式 (A..5) を適用すると式 (A..6) の結果が得られる．

$$\int_0^R \frac{d}{dr}\left\{ar + a(R-r)\right\} dr = \int_0^R \frac{d}{dr}(aR)dr = 0 \tag{A..6}$$

式 (A..5) に示した判定基準より，式 (A..2) のオペレータでは左右対称なパターン分布からユニークな特徴量を抽出できないことが分かった．一方，式 (A..3) のオペレータについては式 (A..7) の結果が得られ，式 (A..4) のオペレータについては式 (A..8) の結果が得られる．

$$\begin{aligned}&\int_0^R \frac{d}{dr}\left[\exp(-ar) + \exp\left\{-a(R-r)\right\}\right] dr \\ &= \int_0^R a\exp(-ar)\left\{\exp(-aR) - 1\right\} dr \neq 0\end{aligned} \tag{A..7}$$

$$\begin{aligned}&\int_0^R \frac{d}{dr}\left[\exp(-ar^2) + \exp\left\{-a(R-r)^2\right\}\right] dr \\ &= \int_0^R 2a\exp(-ar^2)\left[\exp\left\{aR(2r-R)\right\}(R-r) - r\right] dr \neq 0\end{aligned} \tag{A..8}$$

以上より，式 (A..2) は分布に対称性を有しているために形状特徴抽出オペレータとしては不適であるが，式 (A..7) と式 (A..8) の計算結果より，式 (A..3) と式 (A..4) は，オペレータとして使用できることが分かった．本研究では，低域通過フィルタ特性を持たせることが可能なガウシアンオペレータを使用することにする．

A..2 密度係数について

式 (A..1) において式 (A..9) により表される部分を密度係数と呼んでいる．

$$d(\theta) = \frac{R}{\sum_{r=0}^{R} p(r)} \tag{A..9}$$

密度係数を掛け合わせる理由としては，パターン分布区間 $[0, R]$ において，畳み込み演算の際の有効加算回数を R に統一することが挙げられる．有効加算回数を統一することにより，パターン分布における有効要素数に依存しない特徴量を抽出することが可能となる．

A..3 ガウス分布の勾配係数 a の決定方法について

A..1節において，ガウシアンオペレータを適用することにより，対称性を持つようなパターン分布でもユニークな特徴量を抽出することが可能であることが分かった．しかしながら，得られる特徴量のユニークさの程度については，勾配係数 a に依存するために，目的に応じて勾配係数を決定する必要がある．図 A..1 に勾配係数 a を 1 〜 9 まで変化させた際に得られるガウシアンオペレータとその一次導関数を示す．なお，オペレータとして実際に用いるガウス分布は，対象図形の外側輪郭の特徴に関して一層敏感となるように分布の左右を反転させて用いる．式 (A..10) により表される対称性評価関数において r を 0 〜 0.5 まで変化させた場合のグラフを図 A..2 に示す．

$$\begin{aligned} &\frac{d}{dr}\left[\exp(-ar^2)+\exp\left\{-a(R-r)^2\right\}\right] \\ &\quad = \quad 2a\exp(-ar^2)\left[\exp\left\{aR(2r-R)\right\}(R-r)-r\right] \end{aligned} \tag{A..10}$$

図 A..2 に示すように形状特徴量の外側輪郭依存性とユニークさ（uniqueness）

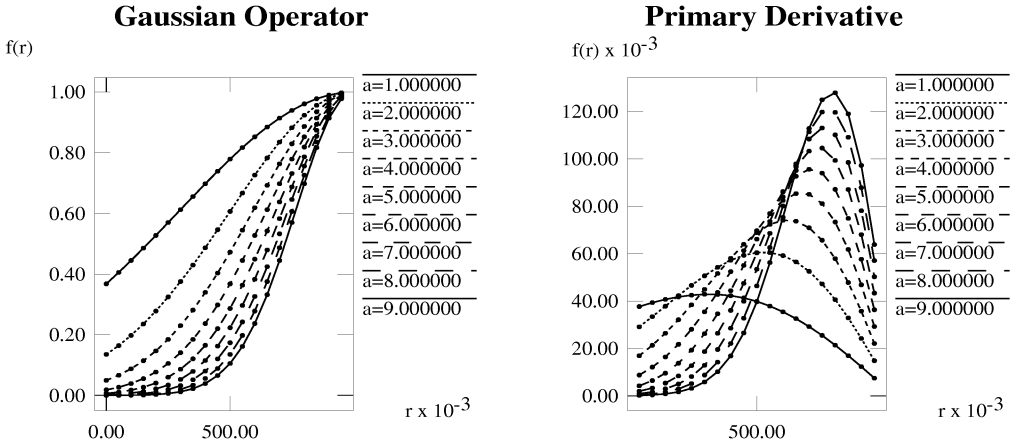


図 A..1 ガウシアンオペレータ（左）とその一次導関数（右）

は勾配係数 a に依存しそれぞれがトレードオフの関係にあることが分かる．以上のことを考慮し，本論文の評価実験においては，勾配係数 a を 5.0 に設定して実験を行った．

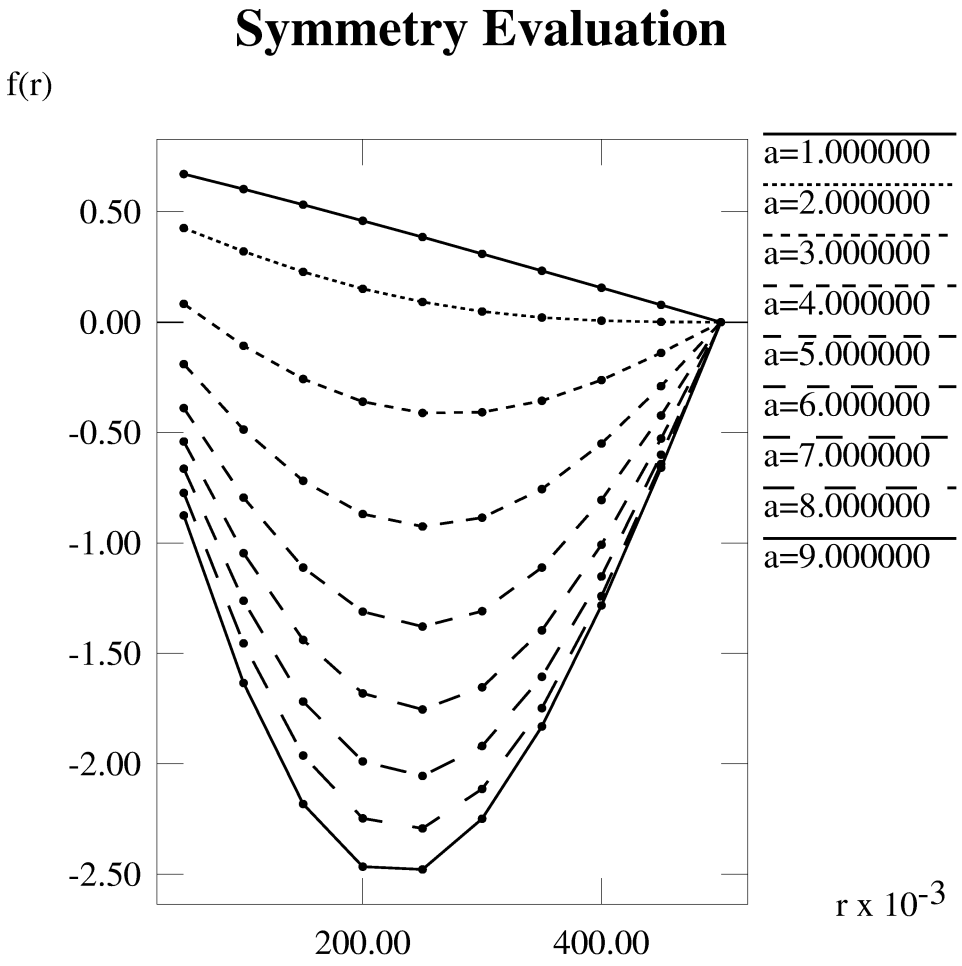


図 A..2 ガウシアンオペレータの対称性評価

B. 評価実験用身振り画像

身振り画像に関しては，現在，画像処理アルゴリズムの客観的評価を行うための SIDBA のような標準画像ライブラリは存在しない．従って，現状では任意で実験に利用する身振りを選定せざるを得ない．本論文では表 B..1 に示す身振りを採用した．

表 B..1 評価実験で用いる身振り

身振り属性	身振り動作名	省略表現
標識（コード）	「バイバイ」	(G-A)
標識（コード）	「グッパ」	(G-B)
例示子（模倣）	「鳥のまね」	(G-C)
情感表示	「バンザイ」	(G-D)
調整子	「聞き返し」	(G-E)
環境適応子	「腕組み」	(G-F)
オブジェクト適応子	「マウス操作」	(G-G)

各身振り属性の簡単な説明を以下に示すが，詳細は文献 [5] を参照されたい．標識身振りとは，明示的情報の伝達を目的とした身振りであり特定の集団ごとに発達し，その成員すべてがそのコードを知っていることを期待される身振りである．例示子身振りとは，対象の形や大きさを空間に描いたり，出来事のリズムを示す動作などに相当するもので会話内容の強調や補足を行うものである．情感表示身振りとは，情感表示の基本的な場は表情であるが，表情に出る感情を補足する場合にとられる二次的な身体動作である．調整子身振りとは，話す順番を決定

したり，発話権のやりとりを制御したり，会話の流れを円滑にする機能をもつ身体動作である．環境適応子身振りとは，身体的要求を満たしたり，情緒を管理したりといった，状況や環境に適応するための身体動作である．オブジェクト適応子とは，道具や機械を使用するために学習されるものであるが，関連する情緒や構えが刺激されると，会話中であっても生じることがある身体動作である．各身振りに対応するスナップショット画像を図 B..1 から図 B..18 に示す.

各被験者の身振りサンプルを収集する際には，例えば,「バイバイの身振りをしてください」などといったキーワードにより身振り動作を求めた．この条件設定により，被験者それぞれが有している身振り概念に対する特徴的動作を獲得することができると考えられる．なお，提案手法が背景や照明に関して頑健であることを確認するために，身振り画像の撮影時には特別な背景や照明は利用していない.

＃１　＃２　＃３　＃４

＃５　＃６　＃７　＃８

図 B..1　バイバイ（標識身振り）

＃１　＃２　＃３　＃４

＃５　＃６　＃７　＃８

図 B..2　グッパ（標識身振り）

＃１　＃２　＃３　＃４

＃５　＃６　＃７　＃８

図 B..3　鳥のまね（例示子身振り）

＃１　＃２　＃３　＃４

＃５　＃６　＃７　＃８

図 B..4　バンザイ（情感表示身振り）

＃１　＃２　＃３　＃４

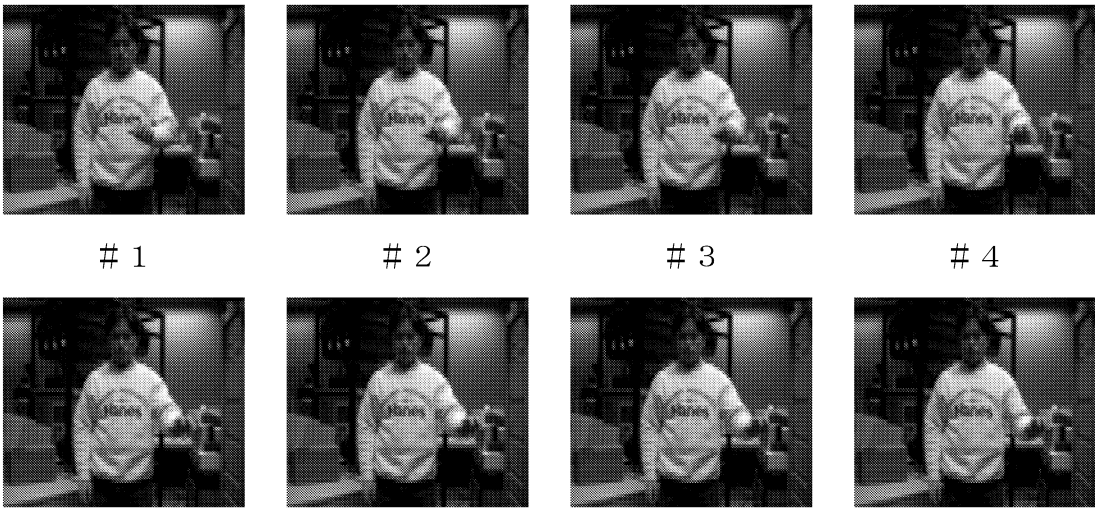
＃５　＃６　＃７　＃８

図 B..5　聞き返し（調整子身振り）

＃１　＃２　＃３　＃４

＃５　＃６　＃７　＃８

図 B..6　腕組み（環境適応子身振り）



＃１　＃２　＃３　＃４

＃５　＃６　＃７　＃８

図 B..7　マウス操作（オブジェクト適応子身振り）


「バイバイ」


「グッパ」


「鳥のまね」


「バンザイ」


「聞き返し」


「腕組」


「マウス操作」

図 B..8 被験者A


「バイバイ」


「グッパ」


「鳥のまね」


「バンザイ」


「聞き返し」


「腕組」


「マウス操作」

図 B..9 被験者B

「バイバイ」

「グッパ」

「鳥のまね」

「バンザイ」

「聞き返し」

「腕組」

「マウス操作」

図 B..10　被験者C

「バイバイ」

「グッパ」

「鳥のまね」

「バンザイ」

「聞き返し」

「腕組」

「マウス操作」

図 B..11　被験者D


「バイバイ」


「グッパ」


「鳥のまね」


「バンザイ」


「聞き返し」


「腕組」


「マウス操作」

図 B..12 被験者 E


「バイバイ」


「グッパ」


「鳥のまね」


「バンザイ」


「聞き返し」


「腕組」


「マウス操作」

図 B..13 被験者 F

「バイバイ」

「グッパ」

「鳥のまね」

「バンザイ」

「聞き返し」

「腕組」

「マウス操作」

図 B..14 被験者G

「バイバイ」

「グッパ」

「鳥のまね」

「バンザイ」

「聞き返し」

「腕組」

「マウス操作」

図 B..15 被験者H

「バイバイ」

「グッパ」

「鳥のまね」

「バンザイ」

「聞き返し」

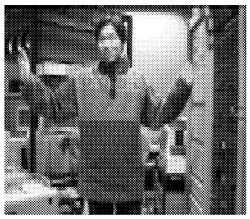
「腕組」

「マウス操作」

図 B..16 被験者 I

「バイバイ」

「グッパ」

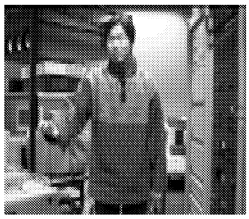


「鳥のまね」

「バンザイ」

「聞き返し」

「腕組」



「マウス操作」

図 B..17 被験者 J

「バイバイ」

「グッパ」

「鳥のまね」

「バンザイ」

「聞き返し」

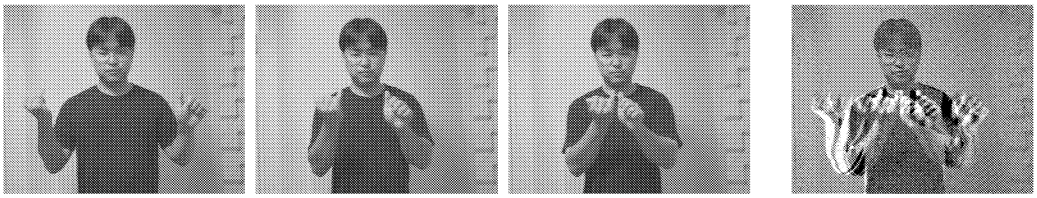
「腕組」

「マウス操作」

図 B..18 被験者K

C. 手話画像データベース収録画像

文献 [49] を参考にして６４種類の手話動作を撮影し手話画像データベースを作成した．手話画像データベースに登録されている画像の総数は８６４枚であり，１手話動作は平均１４枚の濃淡画像により構成されている．図 C..1 から図 C..17 に６４種類の手話動作のスナップショット画像とその軌跡画像を示す．なお，本画像データベースは，厳密な手話画像データベースの構築を目的として作成されたのではなく，同一人物の多様な動作の検索を行うシステムの評価を目的として作成されたことを付記しておく．



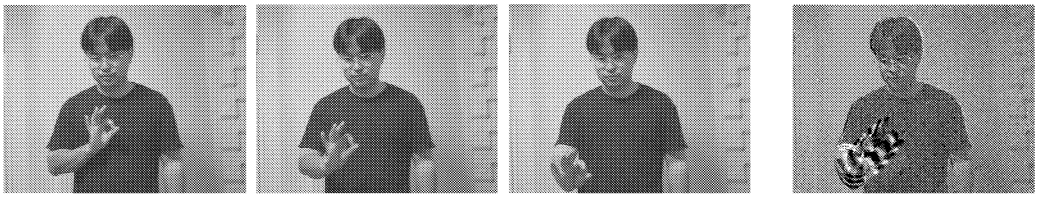
開始　　途中　　終了　　軌跡

「結婚」

開始　　途中　　終了　　軌跡

「ダメ」

図 C..1　収録手話画像（１）



開始　　途中　　終了　　軌跡

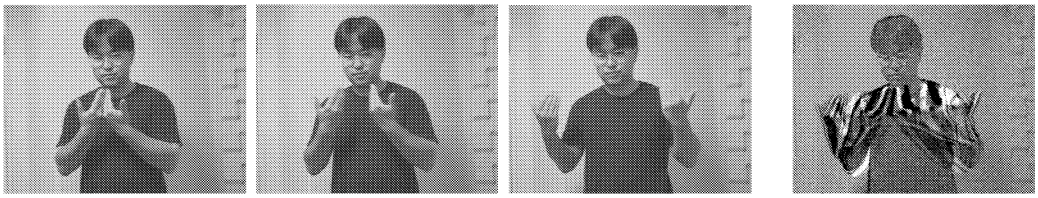
「ＯＫ」

開始　　途中　　終了　　軌跡

「おめでとう」

開始　　途中　　終了　　軌跡

「バイバイ」



開始　　途中　　終了　　軌跡

「ご無沙汰（しています)」

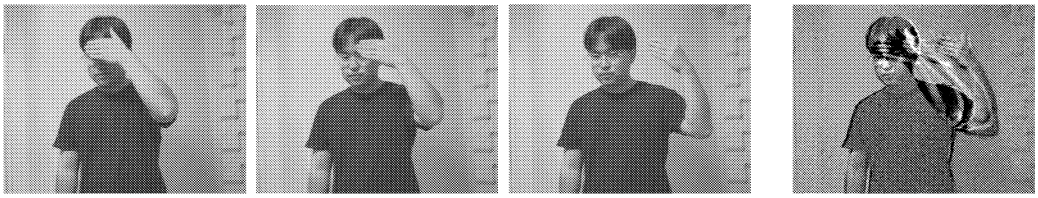
図 C..2　収録手話画像（２）

開始　途中　終了　軌跡

「生まれる」

開始　途中　終了　軌跡

「姉」



開始　途中　終了　軌跡

「青年」

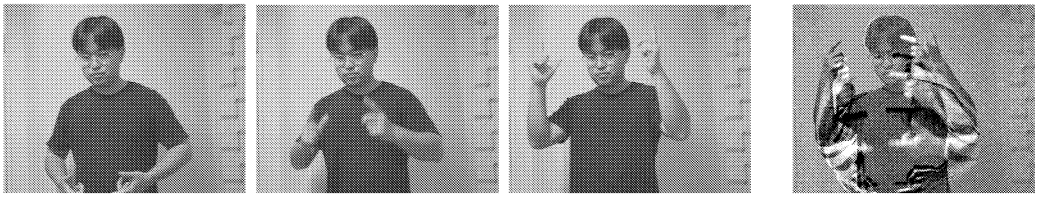
開始　途中　終了　軌跡

「生活」

図 C..3　収録手話画像（３）

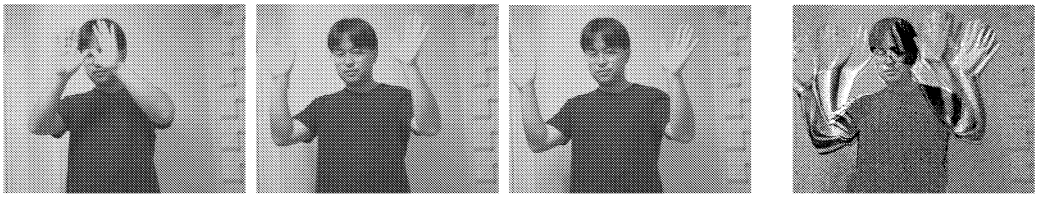
開始　　途中　　終了　　軌跡

「水曜日」



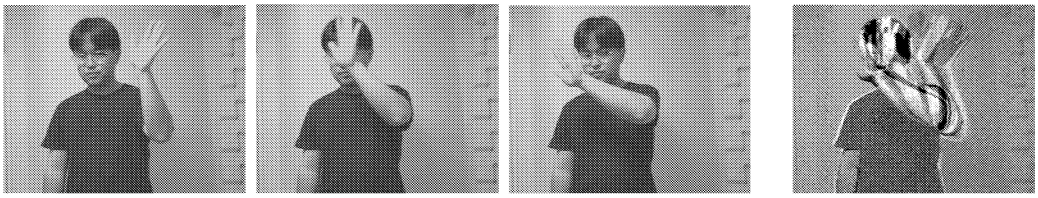
開始　　途中　　終了　　軌跡

「木曜日」



開始　　途中　　終了　　軌跡

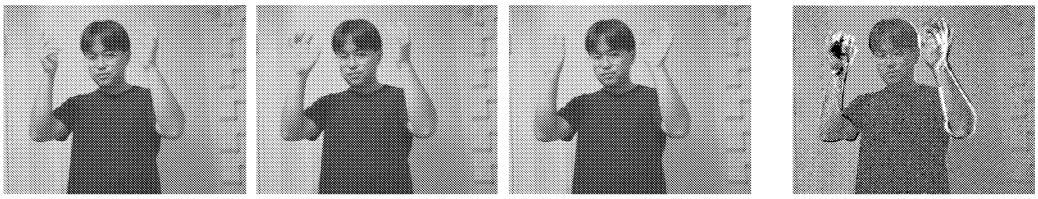
「晴れ」



開始　　途中　　終了　　軌跡

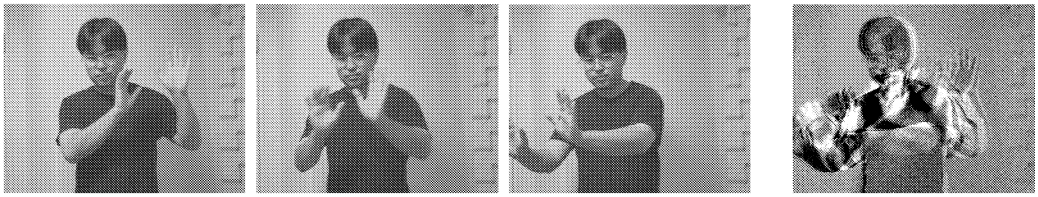
「空」

図 C..4　収録手話画像（４）



開始　途中　終了　軌跡

「雲」

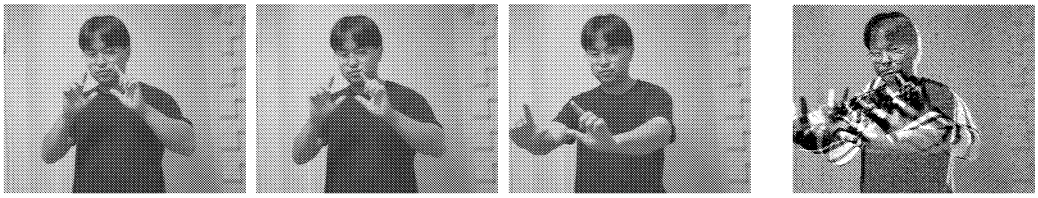


開始　途中　終了　軌跡

「風」

開始　途中　終了　軌跡

「地震」



開始　途中　終了　軌跡

「台風」

図 C..5　収録手話画像（５）

開始　　途中　　終了　　軌跡

「スポーツ」

開始　　途中　　終了　　軌跡

「飲む」

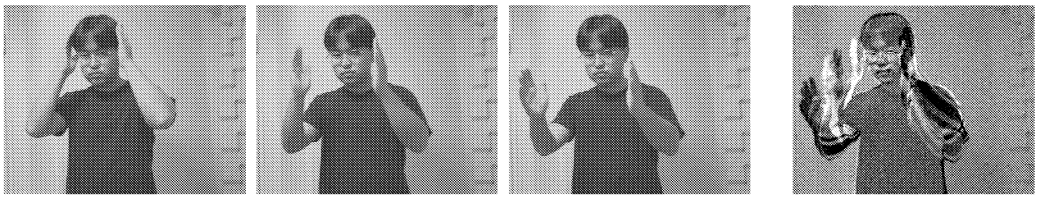
開始　　途中　　終了　　軌跡

「まずい」

開始　　途中　　終了　　軌跡

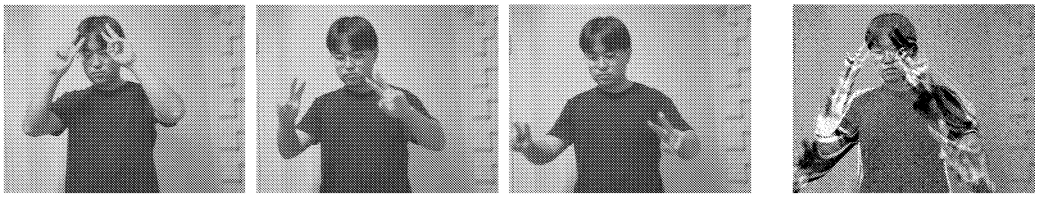
「学校」

図 C..6　収録手話画像（６）



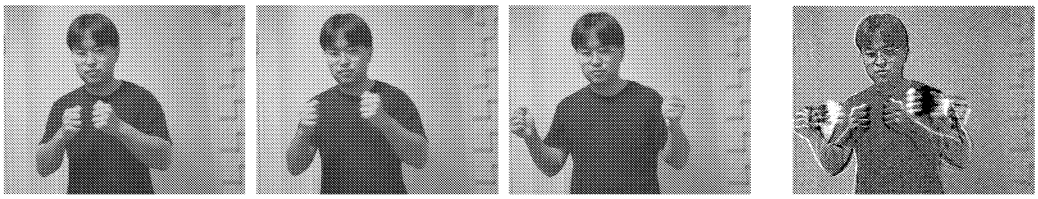
開始　途中　終了　軌跡

「勉強」



開始　途中　終了　軌跡

「音楽」



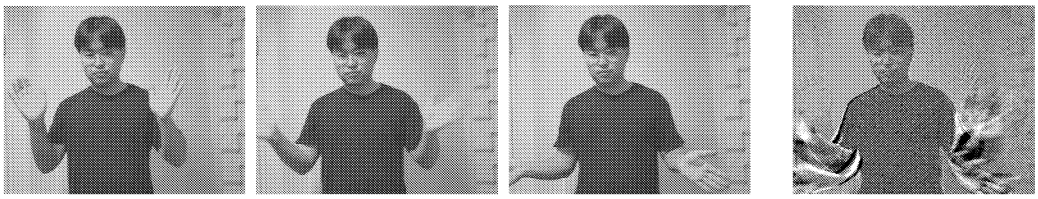
開始　途中　終了　軌跡

「体育」

開始　途中　終了　軌跡

「海」

図 C..7　収録手話画像（７）

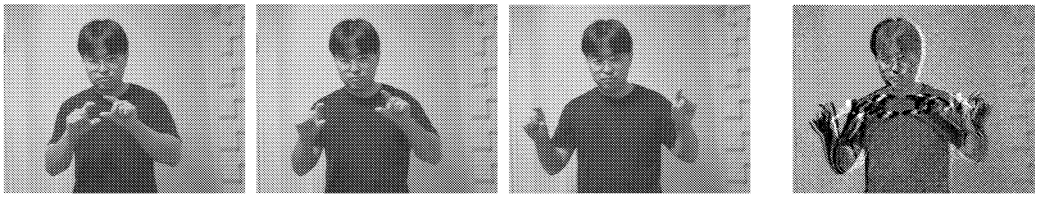


開始　途中　終了　軌跡

「暇」

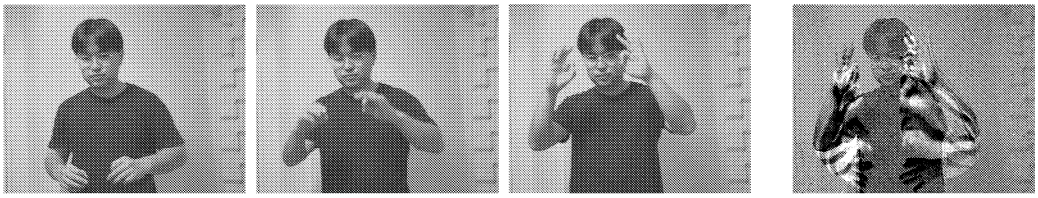
開始　途中　終了　軌跡

「忙しい」



開始　途中　終了　軌跡

「平等」



開始　途中　終了　軌跡

「値上げ」

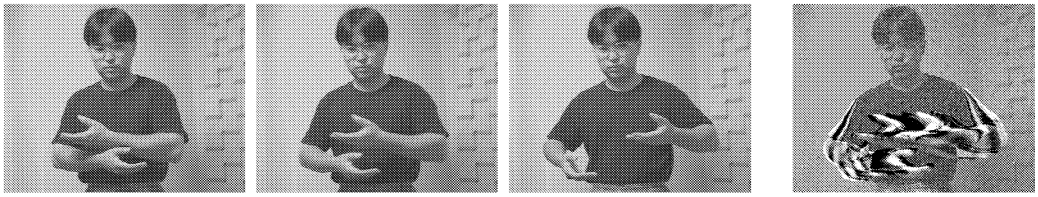
図 C..8　収録手話画像（８）

開始　　途中　　終了　　軌跡

「情報」

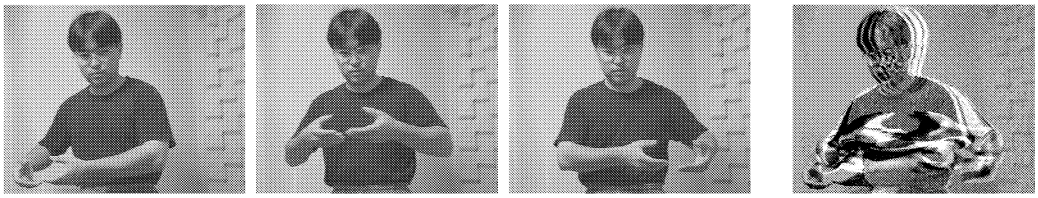
開始　　途中　　終了　　軌跡

「交流（コミュニケーション）」



開始　　途中　　終了　　軌跡

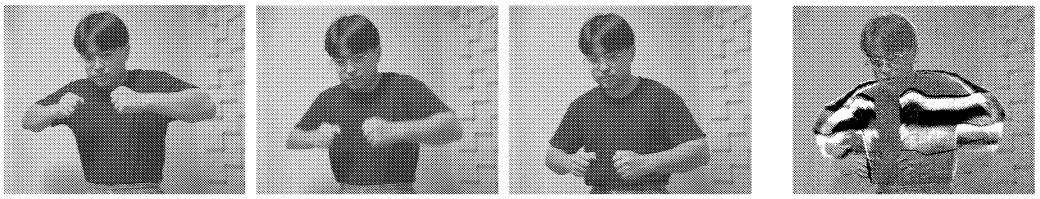
「交際」



開始　　途中　　終了　　軌跡

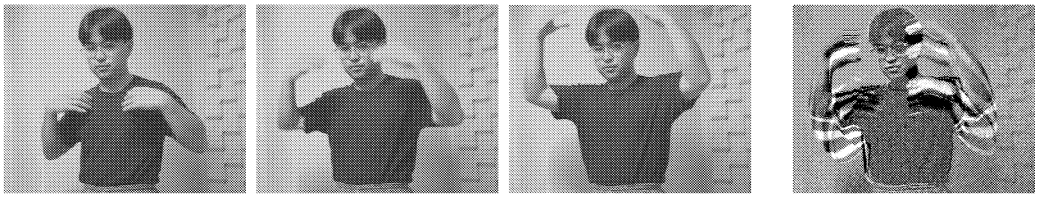
「浮気」

図 C..9　収録手話画像（９）



開始　　途中　　終了　　軌跡

「生きる」

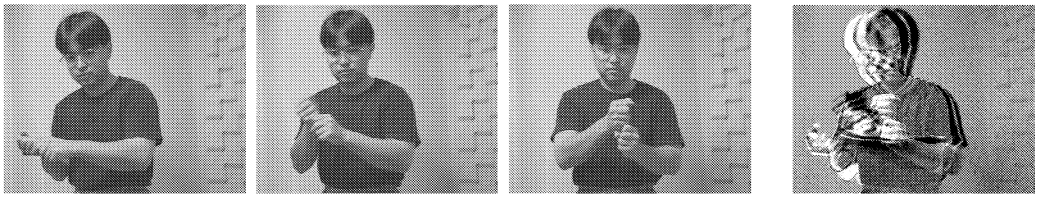


開始　　途中　　終了　　軌跡

「成長」

開始　　途中　　終了　　軌跡

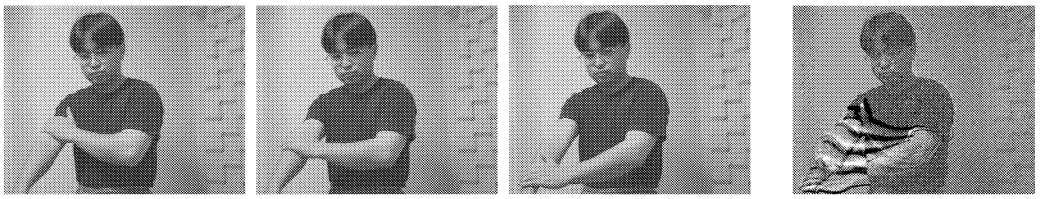
「やっと」



開始　　途中　　終了　　軌跡

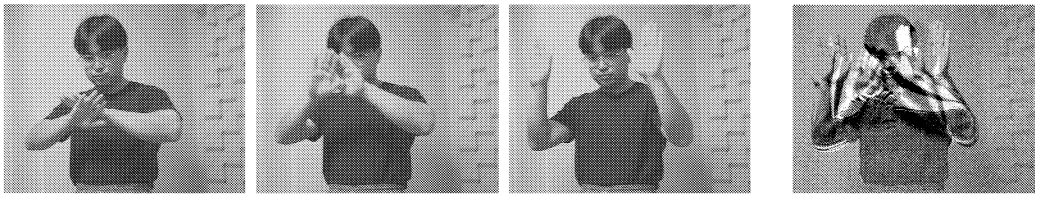
「もう一度」

図 C..10　収録手話画像（１０）



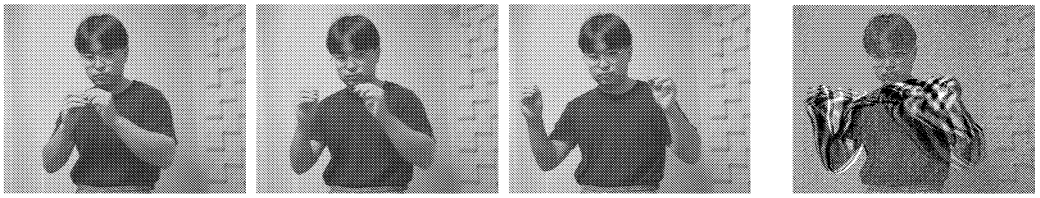
開始　　途中　　終了　　軌跡

「上手」



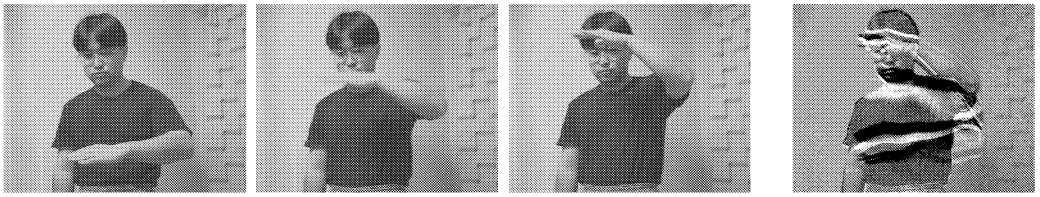
開始　　途中　　終了　　軌跡

「明るい」



開始　　途中　　終了　　軌跡

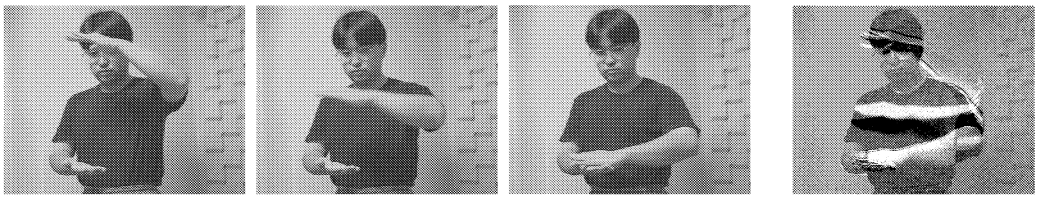
「長い」



開始　　途中　　終了　　軌跡

「高い」

図 C..11　収録手話画像（１１）

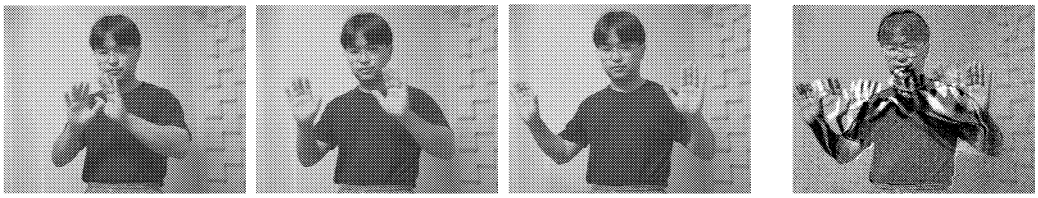


開始　　途中　　終了　　軌跡

「深い」

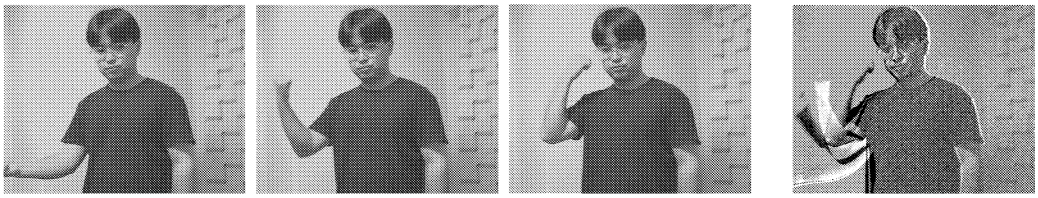
開始　　途中　　終了　　軌跡

「遠い」



開始　　途中　　終了　　軌跡

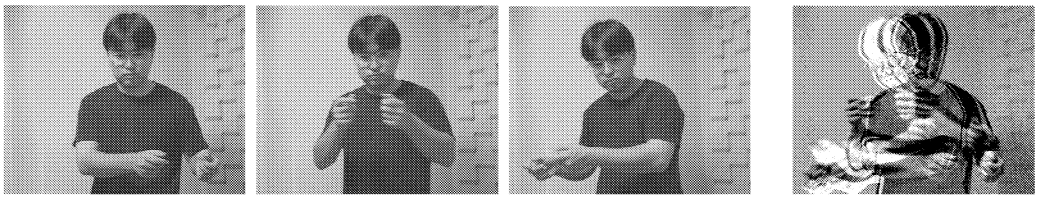
「広い」



開始　　途中　　終了　　軌跡

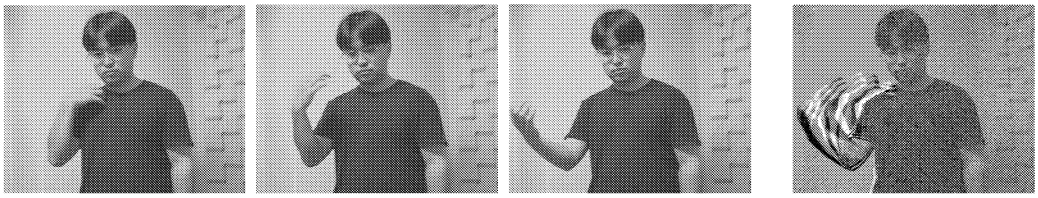
「強い」

図 C..12　収録手話画像（１２）



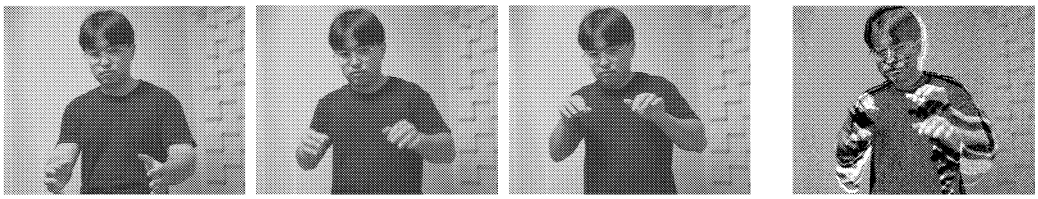
開始　　途中　　終了　　軌跡

「延ばす（延期）」



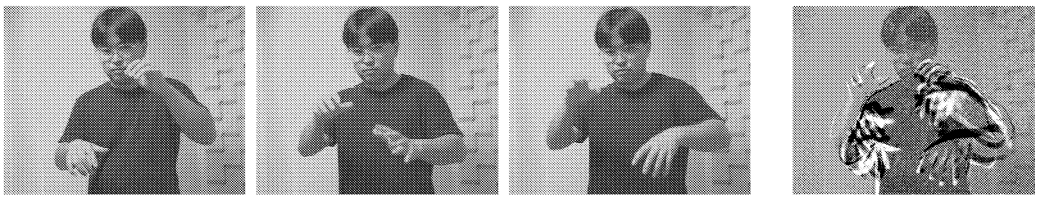
開始　　途中　　終了　　軌跡

「任す」



開始　　途中　　終了　　軌跡

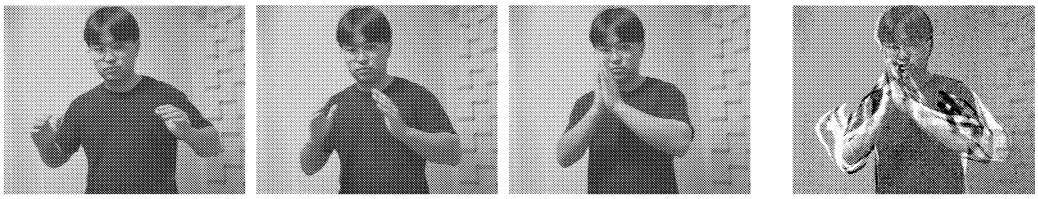
「手を引く」



開始　　途中　　終了　　軌跡

「操る」

図 C..13　収録手話画像（１３）

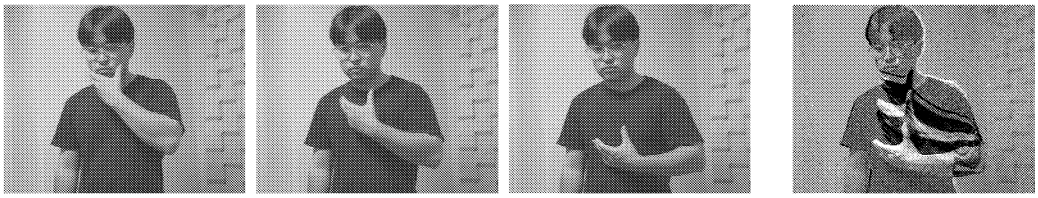


開始　途中　終了　軌跡

「集まる」

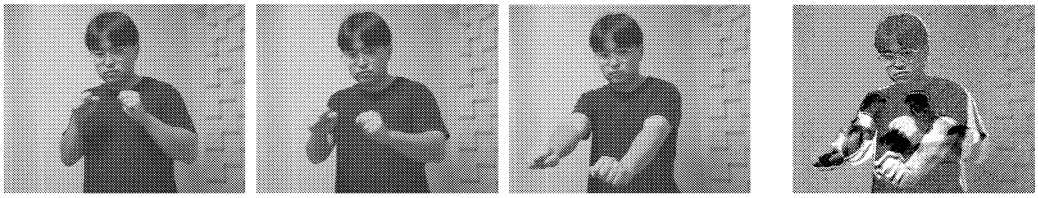
開始　途中　終了　軌跡

「招く」



開始　途中　終了　軌跡

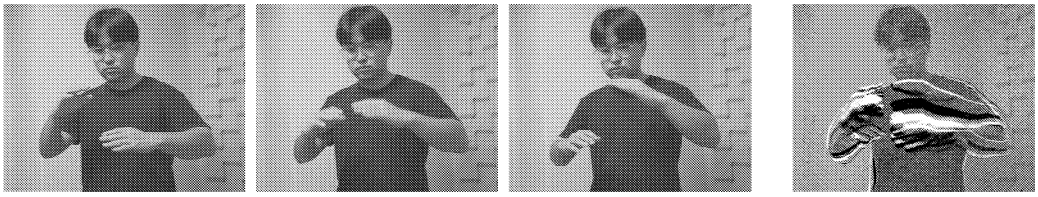
「騙される」



開始　途中　終了　軌跡

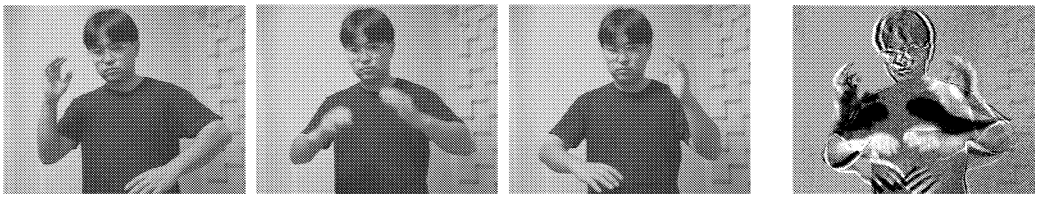
「する」

図 C..14 収録手話画像（１４）



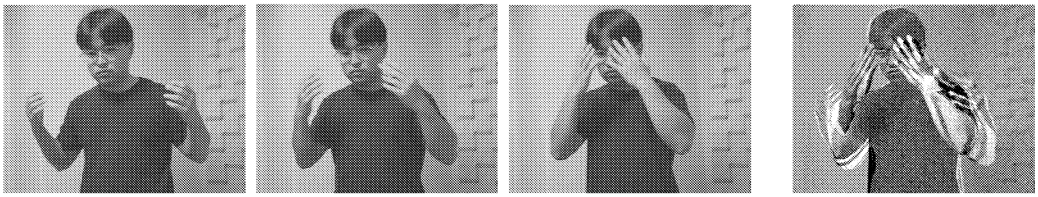
開始　　途中　　終了　　軌跡

「うれしい」



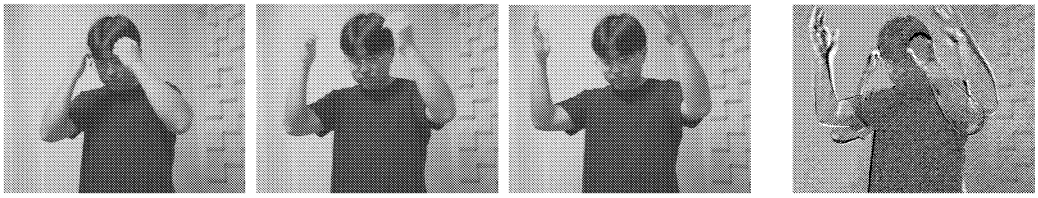
開始　　途中　　終了　　軌跡

「楽しい」



開始　　途中　　終了　　軌跡

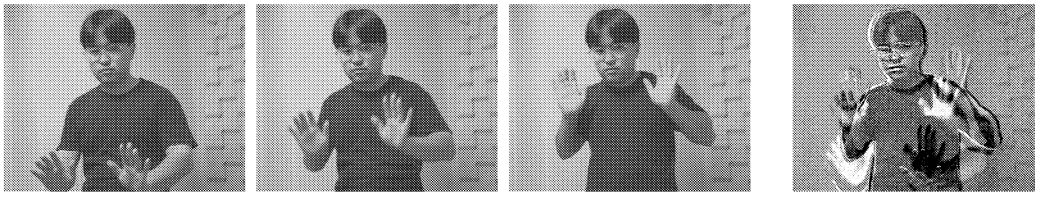
「注目させる」



開始　　途中　　終了　　軌跡

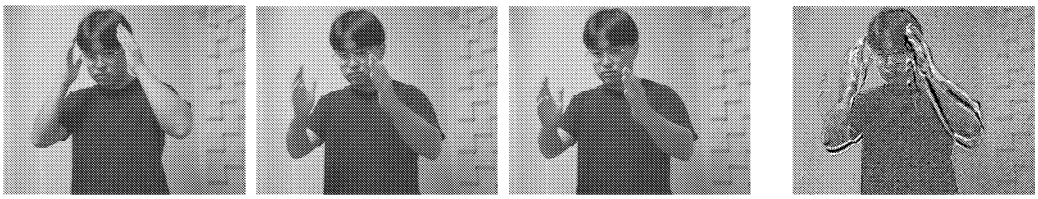
「頭にくる」

図 C..15　収録手話画像（１５）



開始　　途中　　終了　　軌跡

「驚く」

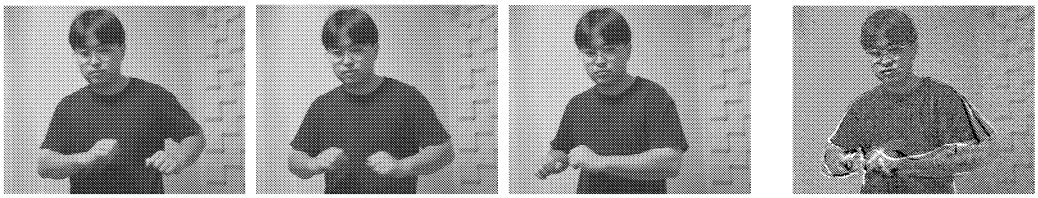


開始　　途中　　終了　　軌跡

「一生懸命」

開始　　途中　　終了　　軌跡

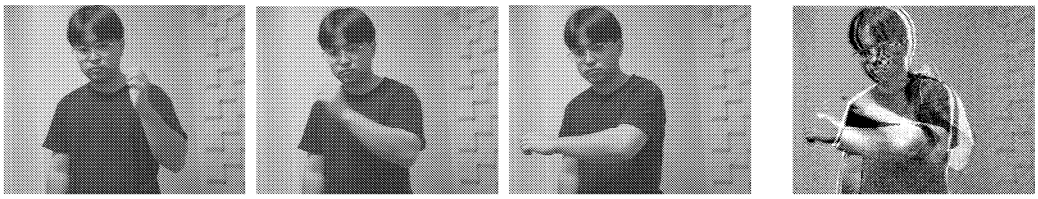
「思う」



開始　　途中　　終了　　軌跡

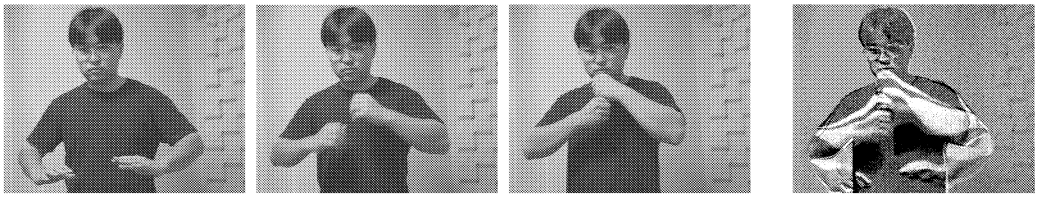
「わがまま」

図 C..16　収録手話画像（１６）



開始　　途中　　終了　　軌跡

「平気」



開始　　途中　　終了　　軌跡

「つまり」

図 C..17　収録手話画像（１７）