注視点の学習と選択制御による身振りの実時間画像認識

桐島　俊之　佐藤　宏介*　千原　國宏**

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

Toshiyuki KIRISHIMA, Kosuke SATO* and Kunihiro CHIHARA**

Currently available gesture recognition techniques and framework share difficulty in recognizing arbitrary person's unspecified gestures in real world and in accomplishing the task at arbitrary speed. In this research, a recognition framework called QVIPS is proposed, which does not assume the types and natures of gestures to be recognized and is suitable for the selective recognition of broader kinds of gesture. In addition, a selective control method for visual interest points is presented, which furnishes with the recognition system self-load monitoring and controlling functionality. To demonstrate the applicability of the proposed methods, a gesture video system and a sign language image database retrieval system are developed. Evaluation results strongly indicate the effectiveness of the proposed methods.

1 まえがき

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。

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。

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。従来、顔表情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。

＊大阪大学大学院 基礎工学研究科
＊＊奈良先端科学技術大学院大学 情報科学研究科

2 本研究における課題

非接触での身振り認識のために、画像処理技術に基づく認識手法がこれまでに数多く提案されている。それらの手法は、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できる。

前者の例としては、円筒モデルなどの幾何プリミティブを入力画像にフイッティングさせる方法や、複数のマーカーを体に取り付けてそれらの位置関係から姿勢などを推定する方法がある。これらの手法はパラメータ推定の収束時間を予測することが困難であり、一定時間内に認識結果を出力することが要求される仮想現実感応用に現状では向いていない。

一方、後者の例としては、入力画像をパターン情報とみなし、それが属するカテゴリーを推定する図形パターン認識技術に基づく手法がある。その代表的手法としては、DP照合法に基づく手法、ファジー連想記憶に基づく手法、HMM（Hidden Markov Mode1）に基づく手法、図形の固有空間内の軌跡の類似性に基づく手法、ニュー

ラルネットワークに基づく手法、更にZemike Momentなどの図形モーメント特徴に基づく手法などがある。事前に認識対象の標準パターン登録作業が必要であるが、画像一枚につき数十ミリ秒から数百ミリ秒で処理されるため、身振り画像の実時間認識に適している。しかし、従来手法は入力画像のクラス推定を主目的とすることが多く、仮想現実感システムへの応用は限定的であった。

身振り画像を認識する場合、処理データ量が膨大となるために、実時間認識するためには、入力画像列からより重要な空間的特徴量を選択的に認識する手法が必要となる。従来、重要な空間的特徴量を選択的に認識する枠組みが考慮されていなかったために、システム開発者の選定した身振りあるいは単一的な身振りをユーザに強制してしまう問題があった。また、従来手法では、身振りが本来備えている多元的な情報（速度、方向、振幅など）の推定は対象としていない。更に、標準パターンの増加に伴う処理フレームレートの低下、仮想現実感システムとの接続による処理フレームレートの低下、OS環境下の他プロセスの影響による処理フレームレートの不安定化、の問題に対応していないため、処理フレームレートを保証することが本質的に困難である。

本論文では、利用者自身がコンピュータに任意の身振りを直接学習させることにより、種々の身振り情報の実時間推定を可能とする多注視点身振り認識法（QVIPS）[(1)]について述べる。続いて、提案手法の仮想現実感応用をより確かなものとする多注視点選択制御法について述べる

3 多注視点身振り認識法

3.1 提案手法の概要

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。

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る。

本研究で提案している多注視点身振り認識法は、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イルタにかけられ、その後、各種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。

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。更に、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、以降、このプロトコルマップに基づいた認識処理が行われる。

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算出される。得られる身振り情報は、標準画像列のフレーム番号に対応する身振り位相値、更に標準画像列と比べた場合の身振りの相対的な速度と振幅である。

3.2 評価実験

3.2.1 実験環境

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160x120で取り込まれ、オンラインで認識される。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実験室内で行った。

3.2.2 身振りプロトコル学習および認識実験

選択的注視が必要となる身振り「バイバイ」（手を振る動作）についてプロトコル学習および認識実験を行った。実験は（1）身振り標準パターンの登録を行う（右手を一回振る）（2）類似身振りによりプロトコル学習を行う（右手と左手をそれぞれ交互に10回ずつ振る）（3）図1に示す類似身振りのテストサンプルを使って認識実験を行う、以上の3段階に分けて行う。プロトコル学習の後、（1）差分画像で最も有効な注視点のみ（2）シルエット画像で最も有効な注視点のみ（3）すべての注視点の中で最も有効な4つの注視点（4）差分画像に関するすべての注視点（5）シルエット画像に関するすべての注視点（6）すべての注視点、以上の6通りの実験条件下で認識実験を行った。ここで（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

図１：類似身振りサンプルのスナップショット（各サンプルとも一枚目の画像のみを示す）

表１：身振り「バイバイ」の認識結果

実験条件	(1)	(2)	(3)	(4)	(5)	(6)
認識率(%)	70	10	100	90	30	100

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。以上の６通りの実験条件下で行った認識実験の結果を表１に示す。条件（３）と（６）の実験条件ですべてのテストサンプルを正しく認識することができた。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（４）に良努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。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（６））が差分画像のみを考慮した場合の結果（条件（１）と（４））を上回っている。この結果は、複数の画像特徴を同時に考慮することにより、単一の画像特徴のみを考慮する場合に比べてより高い認識率が得られることを示している。

4 多注視点選択制御法

4.1 提案手法の概要

身振り認識処理の高速化とその処理フレームレートの安定化は、仮想現実感環境でのインタラクションのリアリティを高める上で極めて重要である[2]が、ソフトウェア処理のみでビデオレートを実現する研究例は少ない。なお、本論文では30[frames/s]をビデオレートと呼ぶ([frames/s]は以降[fps]と表記)。上記の課題に対応するため、図２に示すような、有効注視点の選択制御(Control 1)、パターン照合間隔の選択制御（Control 2)、パターン走査間隔の選択制御（Control 3）の３種類の選択制御系を多注視点身振り認識法に組み込む。

4.2 多注視点選択制御の枠組み

応用システムのみならず認識システムも複数プロセスの一つとして実装されるため、実際の処理負荷の予測は非常に困難である。そこで、図３に示すような処理負荷の大きさを動的に選択するフィードバック制御系を構成することにより、システム全体の応答の安定化を図る。ここでは選択制御を、操作量３変数、制御量１変数のフィードバック制御問題として捉える。具体的には制御量を処理フレームレートx[fps](目標フレームレートv[fps])とし、操作量をパターン走査間隔S_k（特徴画像を走査する際の刻み間隔を指す)、パターン照合間隔RS_k（入力画像から得る特徴パターンと既に登録してある特徴パターンとを照合する際のステップ間隔を指す)、有効注視点数N_{vip}（認識処理で考慮される注視点の数を指す）の３変数とする。

想定される制御対象は、認識モジュール自体の負荷$D_{t1}(t)$、応用システムとのプロセス間通信に伴うネットワーク負荷$D_{t2}(t)$、更に応用システム自体の負荷$D_{t3}(t)$の３種類であるが、本論文では認識モジュールの負荷$D_{t1}(t)$を操作することにより、システム全体での処理フレームレートを目標値へと安定化させる。なお、マルチユーザ・マルチプロセス環境での動作を想定しているので、$D_{t3}(t)$には予測不可能な他ユーザなどのプロセスによる影響も含める。

具体的に、フィードバック制御では、微小操作量を適用した際の処理時間変化を逐次検出し、以降の操作量を決定する。この際、制御偏差e（$=v-x$）が最小となるよう負帰還をかける。ここで、制御状態を単一の指標で

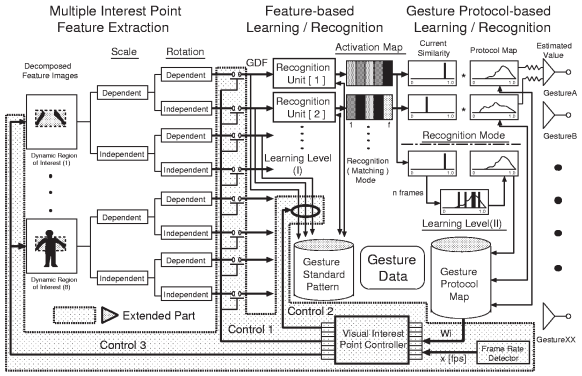


図２：多注視点選択制御に伴う従来手法の拡張

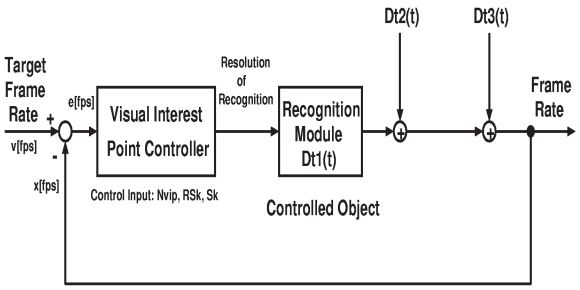


図３：フィードバック制御のブロック図

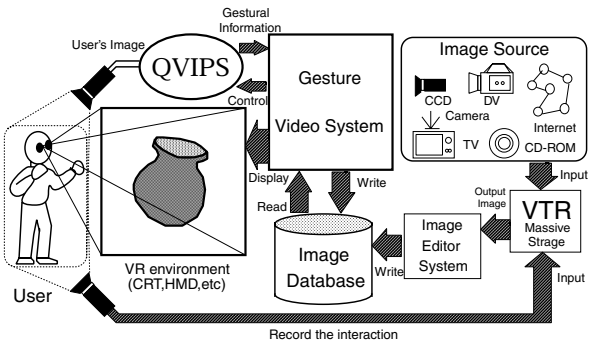
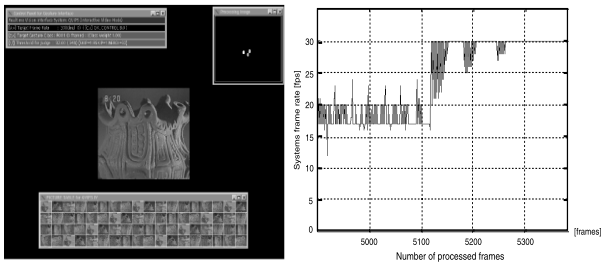


図４：ジェスチャビデオシステムのブロック図



(a)実行画面例　　(b)システム応答

図５：実験結果

表すために、制御指標（Control Index）Sを式(1)により定義する。

$$S = LN(L - S_k) + N(L - RS_k) + N_{vip} \quad (1)$$

ここでNは最大注視点数、Lは制御時の段階数である。制御指標Sが大きいほど、認識精度を重視した処理が行われる一方、小さいほど、認識精度は相対的に低下するものの高フレームレートでの認識処理が行われる。

4.3 評価実験

4.3.1 実験環境

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（Pentium MMX 266MHz、OS:Linux）上にC言語で実装した。CCDカメラからの画像は画像入力装置（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]のサイズで取り込まれ、オンライン処理される。なお、本実験はすべてマルチユーザモード、X-Window上で行った。

4.3.2 応用システムとの接続実験

提案手法がバーチャルリアリティ応用可能であることを示すために、アプリケーションとしてジェスチャビデオシステムを開発した。図４にジェスチャビデオシステムのブロック図を示し、図5(a)に実行画面の一例を示す。次に、物体を眺める動作について多注視点身振り認識法により学習させた。その後、多注視点選択制御法によりビデオレートでのインタラクションが実現されるまでのシステム応答を図5(b)に示す。

図5(b)に示すように、5120フレーム以前では、仮想現実感システムとの接続による処理フレームレートの低下の問題と、OS環境下の他プロセスの影響による処理フレームレートの不安定化の問題の影響が顕著に現れている。選択制御が開始される5120フレーム以降では、システムの処理フレームレートは目標値である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、提案手法により他プロセスの影響を最小限に抑え、かつ、安定した処理フレームレートが実現できることを示している。

5 手話画像データベース検索への応用

5.1 提案手法の概要

本節では、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[3]のみでは対応が困難である。キーワードによる手話画像データベース検索についての研究は既に数多くなされているが、事前に人為的に付与した属性に基づく検索に限定されてしまう問題がある。

本章では、提案手法を応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。

5.2 検索システムの構成方法

検索システムは、図6に示すように、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を付与することにより構成する。

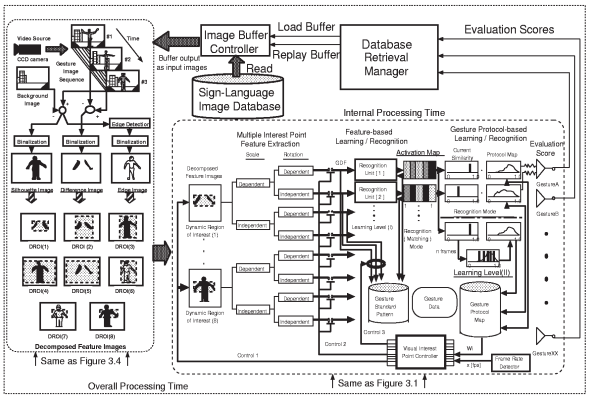


図6:検索システムのブロック図

データベース検索マネージャでは次の［1］～［4］の手続きを実行する。

［1］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。

［2］システム全体での処理時間と内部処理時間を計測する。

［3］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［1］～［3］を繰り返す。

［4］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。

5.3 評価実験

5.3.1 実験環境

手話画像データベースとして文献[4]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、一手話動作は平均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。

提案システムは、手話画像サンプルを学習した後、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、最終的に候補動作名とその評価値を出力する。

5.3.2 任意指定時間での検索実験

本節では、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が可能になることを示す。実験は以下の手順に従って行う。

［1］手話動作として「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、理想的な処理フレームレートを求め、検索システムに設定する。

［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］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。

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(2)により求める。

$$\text{目標フレームレート}=\frac{\text{総検索画像枚数}}{\text{許可される検索時間}} \qquad (2)$$

式(2)により、例えば12秒間での検索（以降、12秒検索と呼ぶ）の場合、目標フレームレートは$\frac{713}{12}\approx 60$に設定すれば良いことが分かる。

図7に12秒検索の結果を示す。プロトコル学習の結果、右手あるいは左手による手話動作が上位候補となっていることが分かる。表2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。

表２：指定時間検索における所要時間とその誤差

許可時間[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

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、必要に応じて認識系を再構成することにより、最善を尽くした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。

図７：12秒検索の結果

6 まとめ

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について述べた。

まず、身振り画像の特徴選択問題を解決するための多注視点身振り認識法について述べた。多注視点身振り認識法では、同一クラス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う。

続いて、身振り画像のサンプリング問題を解決するための多注視点選択制御法について述べた。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る。評価実験では、ビデオレートで任意ビデオ映像とのインタラクションが実現されることを示した。

更に、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。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。

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究して行きたい。

参考文献

⑴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”、電子情報通信学会論文誌D-II, Vol.J81-D-II, No.5, May, 1998, pp.785-794

⑵I.Scott MacKenzie, Colin Ware : “Lag as a Determinant of Human Performance in Interactive Systems”, Proceedings of the Human Factors in Computing Systems INTERCHI’93, 1993, pp.488-493

⑶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

⑷伊藤政雄、竹村茂:“手話入門－ふれあいの言葉－”、1995、廣済堂