解 説

顔とジェスチャの検出および認識

Automatic Extraction and Recognition of Face and Gesture

間瀬健二* *(株) ATR 知能映像通信研究所

Kenji Mase* *ATR Media Integration & Communications Research Laboratories

1. はじめに

人間同士のコミュニケーションにおいて，我々は言葉のほかに，いわゆるノンバーバル言語と呼ばれる，表情，視線，ジェスチャ，姿勢など言葉によらないメディアを利用している．これら顔とジェスチャは，言葉によるコミュニケーションチャネルを補ったり，それだけで理解できるメッセージを形成できることが少なくない．また，非明示的なメッセージや状況を理解するのに，顔やジェスチャが役立つことはよく知られている[1][2].

同じように考えると，将来，人間社会に多くのロボットやコンピュータシステムが導入され，人間の社会生活を支援し活躍してもらうためには，ユーザや作業対象である人間とのコミュニケーションチャネルを太くしておく必要があるだろう．人間の執事のように，1を聞いて10を知って自動的に動くシステムの実現にはまだ時間がかかるかもしれない．しかし，一つの命令を受けたときに，そのとおり実行してよいかどうか，実行するのに不足している情報は何か，などを判断するために，人間のノンバーバル情報を認識理解する機能をシステムが持つことは，好ましいことである．また，騒音のあるような環境ではノンバーバルな命令しか利用できないこともあるだろう．

そのような考えのもとに，すでにロボットビジョンシステムを搭載したロボットシステムの研究や実用例が多く報告されており，人間を対象として回避や追跡などを行うシステムの報告もすでにある（例えば，MIT AI研究所のBrooksらのグループのロボットたち）．監視やサーベイランスのためにも人物画像処理が使われる[3]．また，ゲームやテレビ番組のアニメーション制作のためのジェスチャ計測システム[4]や，VRやマルチメディア・エンタテイメント分野のインタフェースとしての認識システム[5]～[7]の開発なども盛んである．人間という対象が身近である分，その応用分野は非常に広範囲である．

本解説では，ロボット応用に限らない，一般的な顔とジェスチャ認識の基本的手法と最新動向についてまとめる．しかしながら応用面から考えて，非接触処理が可能な画像センシングを前提とした分野に限る．なお，顔の認識については赤松によるサーベイ[8]が，基本的な手法の解説と豊富な参考文献リストにより良くまとめられている．さらに，顔画像の表示やインタフェース応用についてのサーベイ[9]も参考になろう．また，表情認識についても，筆者がまとめた小解説[10]があるので，本稿では，基本的手法については，ジェスチャ認識について重点的にまとめ，顔と表情については簡単に述べることにする．最新の技術動向については，先日開催されたIEEE FG'98（顔とジェスチャの自動認識国際会議）の論文を解説のなかで引用をしながら，トピックを指摘することを試みる．

2. ジェスチャ認識の基本的手法

ジェスチャの認識といっても対象となる身体の部位は，頭，手，腕，上半身，全身などたくさんあり，それぞれその部位の特徴や応用によって様々な手法がとられる．アプリケーションの立場からみると，人間の体から表1のようなメッセージや状況を認識するための手法の開発が行われている[11]．類似の分類は心理学や動作学の分野でも行われており，参考になることが多い[2].

ジェスチャ認識と呼ばれる処理は，大きく以下のような過程に分割することができる．

(1) 三次元モデルの当てはめによる姿勢または連続姿勢の記述
(2) 動作画像からの動物体の時系列特徴抽出
(3) 時系列特徴や記述の，解釈や登録辞書とのマッチング

ここで，(1)と(2)は，動作抽出過程に注目して，その処理をトップダウンに行うかボトムアップに行うかという違いにより分類している．(3)は抽出された動作の識別や認識を行う上位の過程に注目している．研究の多くはこれらのうちどれかに焦点をあてたり，これらを組み合わせて動作するシステムを構築している．

例えば，いま対象とするシステムが必要としている情報が，指先の指示によるポインタ情報だけであれば，指先と

原稿受付 1998年6月10日
キーワード：gesture, face, recognition, image processing
*〒619-0288 京都府相楽郡精華町光台2-2
*Soraku-gun, Kyoto

表1 ジェスチャ認識の対象となる動作と利用シーン

ジェスチャ分類	機能の例	動作の例	利用シーン
指示	身体による空間中の場所指定	指さし，視線，頭の向き	コマンド
位置	身体の空間中の場所	歩く，ユーザと物体の位置関係	状況理解，画像検索キー
姿勢	身体の物体に相対する位置，姿勢	すわる，ユーザの状態検出	状況理解，画像検索キー，計測
操作	物体の操作	機械や計算機操作状態の識別	コマンド，記述，監視
例示	言葉の補足	形状の例示，量や大きさの比喩的例示	コマンド
表象	言語の代りに動作による表出	手話，はい・いいえ	コマンド，メディア変換
情動	感情の表出	表情，力こぶし	インタフェース，監視，診断
身体	生理的動作	頭を掻く，腕を組む	インタフェース，診断
規制	会話円滑化のためのリズム調整	うなずき，手の振り	インタフェース

参照点の三次元位置の特徴量の連続抽出が行われればよい．しかし，もし抽出の信頼性向上や高速化を図りたければ，後述するように，その特徴の時系列を利用して，予測や検証の処理を追加することを考えなければならない．

また，ジェスチャ認識研究は，しばしば実時間性が問われる．これまでは計算能力不足のため，オフライン処理でも利用できる行動分析や動作測定などの分野に限られていた．あるいはインタフェース応用のための研究では実時間性が必須であるため，様々な制約条件を課してもなお，ごく簡単なジェスチャの認識しかできなかった[12]．最近になってコンピュータのビデオ処理能力†が総合的に向上して，少しずつ複雑な条件や対象でも実時間処理が可能となってきている[13]．

2.1 人物の背景からの切り出しと追跡

ジェスチャ認識は，対象が人物のため，背景や環境からの抽出を行うのが普通である．すなわち，一般的な処理手順は図1のように，(i) 人物（の部位）の背景からの切り出しと追跡，を行った後，前述したような，(ii–a) 三次元モデルの当てはめによる姿勢または連続姿勢の記述を行うか，(ii–b) 動作画像からの動物体の時系列特徴抽出，を行い，その時系列特徴や記述を得たあと，(iii) 時系列特徴や記述の解釈や登録辞書とのマッチングをしてジェスチャ認識に至る．

では，処理対象となる人物あるいは部位を背景から切り出すには，どうするか．多くは，(1) 最初から背景は変化しない，(2) 部位の形状あるいはパターンが既知でモデル化できる，(3) 対象部分の動きは背景に対して十分大きい，といういずれかの仮定で処理をしている．

2.1.1 変化しない背景からの切り出し

まず，背景が変化しない場合には，最初に背景パターンを覚えさせ，次々に入力される画像から各画素ごとに差分をとり，前景の対象部分を浮かび出させる．単純なようであるが，実際には背景はいろいろな要因で微妙に変化する．太陽の動きや雲，蛍光灯のちらつきをはじめとする照明の変化，人物と背景との相互反射，人物が落とす影，背景にある物体の移動など様々である．また対象部分が記憶した背景と同じ色の場合には，背景として抜けてしまう．これらの問題を完全に解決するのは困難である．しかし，(a) 背景の変動はスムーズである，(b) 統計的には背景の変動はごくわずか，(c) 影は色相には大きく影響しない，といった統計的な背景画像モデルを仮定したり，光の反射モデルからの知見を利用することで，かなり対象パターンを抽出することができる．多くの人物追跡プログラムはこの方法を使っており，例えばWrenらのPfinder[13]は，SGI O2で動作する，一人の人物を対象としたジェスチャ認識プログラムであるが，あらかじめ背景画像を取り込み，多次元統計量で背景と対象の色と形状をモデル化して，領域を抽出している．緩やかな環境の変化にはこのモデルは対応で

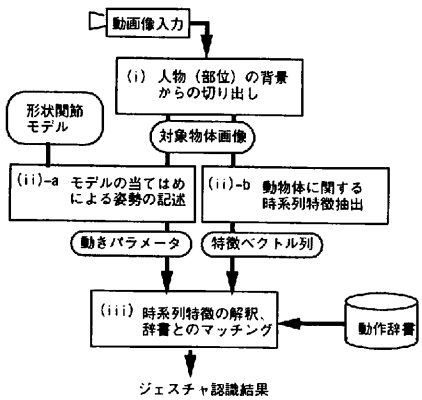

図1 ジェスチャ認識の一般的な処理手順：ステップ (ii) は一方あるいは組み合わせて用いられる

†CPU，メモリアクセス，ビデオメモリ，ビデオバス，A/Dなどが関係する．

きるが，背景にあるものが動いたりすると途端に破綻する．屋外での多人数の検出と追跡をPCで処理することを目標としたW^4 [14]もやはり，背景の学習フェーズがあり，各画素の最大値，最小値，最大フレーム間差分値でモデルを形成し，ある程度動作中に修正している．なお，光源の変動に弱いということは，光源を完全に制御下に置いてしまえば話は簡単である．人工的な近赤外線光源を作りその反射画像（近赤外画像）を用いることが考えられる[12]．

2.1.2 モデル化できる対象の切り出し

対象部分の切り出しには対象の形状や色やパターンのモデルが使われることも多い．対象物体が分からなくとも背景がブルースクリーンのように既知であればそれをモデルとして本来の対象を抜き出すこともできる．可視光ではなく赤外領域の熱画像を使うのもこれに近い[5]．単純な背景差分では対象部分が抜けてしまうことの対策として，形状モデルを利用して補完することが行われる．また，背景の差分を用いず，対象のモデルを積極的に用いて対象部分を抽出する方法もある[15]．しかしながら，モデルとのマッチングコストは高く実時間性に難点があり，初期フレームをマニュアルで指定することで問題を軽減するなどしている．なお，対象を形状とせず，手，顔など露出した肌の色情報を使ったり，色付きマーカなど補助手段を装着，接着して，その動きを検出して対象部分の動きとする方法も高速かつ安定した手法としてしばしば採用される[16]．

2.1.3 動きの大きい対象の切り出し

背景に比べ人物の動きが十分大きいとき，フレーム間の画像差分の絶対値をとることによって，人物の動いている領域をマスクとして抽出することができる．厳密には，2フレームにおける見かけの動領域の和集合が得られる．輝度差分がなければ見かけの動きはない．このマスクと原画像のテクスチャや領域セグメンテーションを使って原画像から動領域を抽出する．さらに進めて，Cutlerら[17]は，Pentium II（266[MHz]）を使って，オプティカルフローの実時間計算を行い，手のジェスチャ領域抽出が可能であると報告している．

なお，これらの処理でリアルタイムと呼ぶのは，20〜30[Hz]のフレームレートであり，解像度はビデオの約1/2から1/4（320×240から160×120）程度である．

2.2 モデルの当てはめと特徴抽出

背景から対象となる人物のみを抽出した対象画像を得られたら，モデルの当てはめをしたり特徴パラメータへの変換をする．上記の例のように，背景から抽出する段階からモデルを利用することもあるが，ここでは説明のため，別のプロセスとしてとらえる．

まずトップダウンのモデルベース的な手法では，対象となる部位をCADの関節モデルのように，円筒，ブロックあるいはスティック部品のリンクで記述しておいて，得られた対象画像に当てはめる．例えば手のモデルは手のひらと1〜2個の関節をもつ指がつながっているように記述できる．当てはめにより各関節の角度が決定され手形状が抽出できたことになる．当てはめには，領域の輝度特徴や，エッジ特徴を手がかりとし，また関節モデルの運動上の制約（順キネマティクスや逆キネマティクス）を使いながら，繰り返し法で誤差を最小化する手法が用いられる．モデルの自由度が高くなると，当てはめの空間が広がり曖昧さが増すため，目的に応じたモデルの精密さが必要である．例えば，全身の三次元モデルは，部品10，自由度22[15]で，手の三次元モデルは部品16，自由度15で記述している[18]例がある．また，前述の文献[13][14]なども全身の二次元モデルを使っており，部品数がそれぞれ6および8という単純さである．

ボトムアップ的な特徴ベースの手法では，対象画像が得られたら，次元を落して冗長性を減らすような，何らかの特徴空間にマッピングする．例えば，文字認識でよく使われる，画像をメッシュ分割し特徴ベクトルを作る方法[19]，固有空間法で全パターンを最も良く記述する部分空間を作る[20]方法がある．新しい試みとして音声特徴で有名なPARCOR係数[21]を画像に適用して局所自己相関特徴を利用する手法が提案されたりしている．

2.3 解釈とマッチング

特徴ベクトル列が得られたら，いよいよそのジェスチャを解釈したり認識して記号化する．最も単純な方法は登録してあるパターンとのマッチングである．しかしながらジェスチャは人間がするため，そのスピードやタイミングが登録パターンと一致することはめったにない．そこで，音声認識で用いられた，DP（Dynamic Programming）マッチングやHMM（Hidden Markov Model）などがジェスチャや表情認識の分野でも注目されている[19][21][22]．

記号化された動作パラメータ列の解釈については，パラメータ空間で軌跡を描いて得られる位相軌道で表現することで動作の識別が可能である[23]．

3. 顔と表情

顔の認識についても，正面の静止画像を対象とした研究から始まって，現在は表情や動きのある顔の処理が注目されている[8]．むしろ1枚の画像で苦労せずに，複数の画像から認識しやすい画像を選ぶことが可能となったと考えるべきであろう．顔の認識の場合も，背景からの抽出，造作の検出と位置合わせなどを行って特徴の正規化の準備をしておいてマッチングをするのはジェスチャ認識と基本的な処理方針は同じである．最近は特徴記述とマッチングには，固有空間法をベースとした手法が支配的である（例えば固有顔を使う方法[24]）．顔の認識は米国国防省のFERETプログラム主導で，アルゴリズム競争が行われ，高い認識

率が達成されている．文献[25]は，その報告書的な論文である．

姿勢，照明，表情の変動があるような状況での顔の識別は相変わらず難しいが，Edwardsら[26]は個別性の変動と状況の変動を分離するような最適な追跡手法を提案しており，興味深い．

4. むすび

顔とジェスチャの認識を概観した．プロセッサやメモリ技術の向上により，最近の手法は，処理の複雑化，統合化が行われ，ストレートな処理過程で説明するのは困難になっている．技術解説のために，引用した論文の主旨とは違う箇所に焦点を当てたものもある．お許し願いたい．興味のある方は，ぜひ原著にあたっていただきたい．

ビジョンセンサと処理エレメントを統合したビジョンチップの開発も，期待と発展の余地がある．動画像処理は処理フレームレートが上がれば，もっと処理が単純化できるという，一見矛盾めいた関係がある．Ishikawa [27]の1[ms]クロックで動作する，完全ディジタル処理の試作ビジョンチップが制御するカメラは，単純な処理ながら人間やボールをまったく遅れずに追跡できている．また，Kageら[28]のアナログ処理による人工網膜チップなどもゲーム市場でヒットすれば，さらに性能向上への研究開発に拍車がかかるだろう．近い将来，これらのデバイスで顔とジェスチャ認識手法のパラダイムは大きく変化するかもしれない．

参考文献

[1] V. Bruce: Recognizing Faces. Lawrence Erlbaum Assoc., London, 1988.

[2] マジョリー＝ヴァーガス：非言語コミュニケーション（石丸 正訳），新潮選書，新潮社，1987.

[3] 増田功："セキュリティのための人物像の自動認識"，映像情報メディア学会誌，vol.51, no.8, pp.1154–1158, Aug. 1997.

[4] 福井一夫："モーションキャプチャ"，映像情報メディア学会誌，vol.51, no.8, pp.1120–1123, Aug. 1997.

[5] S. Iwasawa, K. Ebihara, J. Ohya and S. Morishima: "Real-time Human Posture Estimation using Monocular Thermal Images," In Proc. of the Third Int'l Conf. on Automatic Face and Gesture Recognition (Proc. of FG'98), pp.492–497, Nara, Japan, Apr. 1998.

[6] S. Fels and K. Mase: "Iamascope: A Musical Application for Image Processing," In Proc. FG'98, pp.610–615, Nara, Japan, Apr. 1998.

[7] R. Kadobayashi, K. Nishimoto and K. Mase: "Design and Evaluation of Gesture Interface of an Immersive Walk-through Application for Exploring Cyberspace," In Proc. of FG'98, pp.534–539, Nara, Japan, Apr. 1998.

[8] 赤松茂："コンピュータによる顔の認識"，電子情報通信学会論文誌，vol.J80-D-II, no.8, pp.2031–2046, 1997.

[9] 長谷川修，森島繁生，金子正秀："「顔」の情報処理"，電子情報通信学会論文誌，vol.J80-D-II, no.8, pp.2047–2065, 1997.

[10] 間瀬健二："表情の自動認識"，映像情報メディア学会誌，vol.51, no.8, pp.1136–1139, Aug. 1997.

[11] 間瀬健二："マルチモーダル・インタフェースのための画像処理"，第2回画像センシングシンポジウム講演集，pp.123–128, June 1996.

[12] M. Fukumoto, K. Mase and Y. Suenaga: "Finger-pointer: Pointing Interface by Image Processing," Comput. & Graphics., vol.18, no.5, pp.633–642, May 1994.

[13] C.R. Wren, A. Azarbayejani, T. Darrell and A. Pentland: "Pfinder: Real-Time Tracking of the Human Body," IEEE trans. PAMI, vol.19, no.7, pp.780–785, July 1997.

[14] I. Haritaoglu, D. Harwood and L.S. Davis: "W^4: Who? When? Where? What? — A Real Time System for Detecting and Tracking People," In Proc. of FG'98, pp.222–227, Nara, Japan, Apr. 1998.

[15] M. Yamamoto, T. Kondo, T. Yamagiwa and J. Yamanaka: "Skill Recognition," In Proc. of FG'98, pp.604–609, Nara, Japan, Apr. 1998.

[16] L. Goncalves, E.Di Bernardo and P. Perona: "Reach Out and Touch Space," In Proc. of FG'98, pp.234–239, Nara, Japan, Apr. 1998.

[17] R. Cutler and M. Turk: "View-based Interpretation of Real-Time Optical Flow for Gesture Recognition," In Proc. of FG'98, pp.416–421, Nara, Japan, Apr. 1998.

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

[19] 大和淳司，倉掛正治，伴野明，石井健一郎："カテゴリー別VQをもちいたHMMによる動作認識法"，電子情報通信学会論文誌，vol.J77-D-II, no.7, pp.1311–1318, July 1994.

[20] T. Watanabe and M. Yachida: "Real Time Gesture Recognition Using Eigenspace From Multi Input Image Sequences," In Proc. of FG'98, pp.428–433, Nara, Japan, Apr. 1998.

[21] T. Kurita and S. Hayamizu: "Gesture Recognition using HLAC Features of PARCOR Images and HMM based Recognizer," In Proc. of FG'98, pp.422–427, Nara, Japan, Apr. 1998.

[22] 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"，電子情報通信学会論文誌，vol.J80-D-II, no.6, pp.1563–1570, June 1997.

[23] L.W. Campbell and A.F. Bobick: "Recognition of Human Body using Phase Space Constraints," In Proc. of 5th ICCV, pp.624–630, 1995.

[24] M. Turk and A. Pentland: "Face Recognition Using Eigenfaces," in Proc. CVPR'91, pp.586–591, 1991.

[25] S.A. Rizvi, P.J. Phillips and H. Moon: "The FERET Verification Testing Protocol for Face Recognition Algorithms," In Proc. of FG'98, pp.48–53, Nara, Japan, Apr. 1998.

[26] G.J. Edwards, C.J. Taylor and T.F. Cootes: "Learning to Identify and Track Faces in Image Sequences," In Proc. of FG'98, pp.260–265, Nara, Japan, Apr. 1998.

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

[28] H. Kage, K. Tanaka and K. Kyuma: "3-D Human Motion Sensing by Artificial Retina Chips," In Proc. of FG'98, pp.522–527, Nara, Japan, Apr. 1998.

間瀬健二（Kenji Mase）

1956年生．1981年名大大学院工学研究科修士（情報）課程修了．同年日本電信電話公社（現NTT）入社．1988～89年米国MITメディア研究所Research Affiliate．現在（株）ATR知能映像通信研究所第二研究室長．CG，画像処理，エージェント技術のヒューマンインタフェース応用研究に従事．人工知能学会1997年度研究奨励賞受賞．博士（工学）．