マルチモーダル・インタフェースのための画像処理
Image Processing for Multi-Modal Interface

間瀬健二
Kenji Mase

ATR 知能映像通信研究所
ATR Media Integration & Communications Research Laboratories
e-mail: mase@mic.atr.co.jp

Abstract

キーボードやマウスといった入力デバイスを手で操作するという従来型のコンピュータとコミュニケーションをはかる方法に対して、音声情報、映像情報、触覚情報など複数の情報源を用いた自然な対話を目指すマルチモーダル・インタフェースの研究が盛んである。人間が多様な感覚器官で情報を入手するように、コンピュータもキーボードやマウスだけでなく目や耳を与えることによって状況を的確に把握できるようになると考えられるからである。本文ではマルチモーダル・インタフェースにおける画像処理の役割を例を示しながら述べ、さらに、インタラクティブでクリエイティブなMICインタフェースへの道を探る。

1 はじめに

キーボードやマウスやボタンのような入力デバイスを手で操作するという従来型のコンピュータ・インタフェースに対して、音声情報、映像情報、触覚情報など複数の情報をセンシングすることによって、自然なインタラクションを可能とするインタフェースの研究が盛んに行われている。あたかも人間同士で対話しながら相手に意図や命令を伝えるかのように、コンピュータや機械を使うことができたらどんなにか便利だろう。

でもほんとうに便利だろうか？いったいどんな状況で機械やそのインタフェースに何を期待するのかよく考える必要がある。たとえば、車で、ある目的地に行くときに、「もうちょっと右」とか「少し戻して」とか教習所の教官よろしく、言葉を使って車に命令したいのだろうか？私たちが期待しているのは、タクシーのドライバのように、「京都駅まで」といえば「新幹線口ですね？」と気をきかせてくれて、安全に早く連れていってくれるようなシステムではないか。すなわち、ハンドルの直接操作を人間的な対話でできるようにするインタフェースが欲しいのではなく、ハンドルやアクセルを制御するプログラムに対してマクロな命令を解釈してくれるインタフェース、ドライバの代りの役目を果たしてくれるインタフェース・エージェント [1] を必要としているのである。[1]

もう少しドライバの反応を調べてみよう。「新幹線口」だと思ったのは、大きな荷物をもっているのを見たか、東京弁で話すのを聞いたからかもしれない。つまり、“誰”が“どこ”でそのメッセージを発したかということが、“どんな内容”か以上に重要 [2] なことがあるし、メッセージの解釈に大きく影響している。

このように、同じ信号でもその状況や他の信号と統合することで別の解釈にもなりうる情報群を処理できるインタフェースが必要であると考えられる。マルチモーダル・インタフェースとは、このような情報群の主たる発信源である人間から、状況に適切なメッセージを抽出して機械との自然なインタラクションを実現してくれるインタフェースである。それが特に威力を発揮するのはこのような情報の集約により、ユーザへ気軽さを提供し解釈の曖昧さを除去する効果 [3] が得られると

[1] 自分でドライブを楽しみたい人にこんなシステムを押し付けるつもりはない。むしろ、危険がないか周りの状況を把握して必要なときに教えてくれるシステムやインテリジェントなナビゲーションシステムを提供すべきであろう [2]。

きである。

マルチモーダル・インタフェースの処理を分類すると、信号のセンシング、単独メッセージの抽出、統合、解釈というステップを踏む。本稿では、主に画像信号を対象として各ステップを概観する。まず動作学などにおけるボディランゲージ（あるいはノンバーバル言語）のマルチモーダル・コミュニケーションにおける役割分析を概観したあと、単独メッセージの抽出から統合解釈において使われている画像処理技術を、具体例を示して現状や課題を解説する。なお、紙面の都合で各画像処理アルゴリズムに関する詳細は省く。[2, 4, 5, 10] の文献などを参照頂きたい。

2 マルチモーダル・インタフェース

最初にマルチモーダル・インタフェースという用語の定義について私なりの考えを示しておこう。類語にマルチメディア・インタフェースがあるが、マルチメディア・インタフェースは単にメディア（音、映像、触覚など）が複数になっているときを表すのに対し、それぞれのメディアがいろいろな形態で使われ情報伝達を行っているときに、マルチモーダル・インタフェースと呼ぶと考えられる。例えば、同じ音でも言葉としての音声、韻律、擬態語、摩擦音や落下音、のように分類するとモーダリティを考えることができる。あるいは人差指を伸ばした動作の映像は、1という数字、物体の指示、口にあてて静かにという命令、など数種類のメッセージを手の同じ映像というメディアから伝達するときにマルチモーダルだということができる。

2.1 人間対人間の場合

人間対人間の場合はインタフェースというより、マルチモーダル・コミュニケーションと呼ぶほうが適切であろう。人間対人間のコミュニケーションにおいてマルチモダリティがいかに使われるか、ノンバーバル言語 (non-verbal language) や動作学 (Kinesics) の研究から学ぶことができる。

人間にはメッセージ発出の仕方に複数のモダリティがあって、人間同士のコミュニケーションがマルチモーダルであることは、前述の例でも明らかである。例えば、P.Ekman は身振り手振りや表情などの動作は次の5つに分類できるといっている。(1) 表象動作 (emblem): メッセージを意図的に伝達するときに使われる。「はい」や「いいえ」の首の動きなど。(2) 身体操作 (body manipulator): 意図的でないメッセージ伝達動作で、おもに自分自身の体に作用する動作。頭を掻く、鉛筆をもてあそぶなど。(3) 例示的動作 (illustrator): ことばによるメッセージを説明・例証する動作。話題の対象を指さしたり、手で山の形を描く動作がこれにあたる。表象動作が動作だけで完結するのに対して、例示的動作はことばが伴う。(4) 情動表出 (emotional expression): 顔の表情など、個人の情緒的な状態や反応を伝える。(5) 規制的動作 (regulator): ことばによるコミュニケーションを監視し規制する動作。注目や相づちなどがある [6, 7]。

また、Cassell[8] によれば、McNeill は発声と共に生じる手のジェスチャをさらに詳しく分析し、(1)Iconic, (2)Metaphoric, (3)Beat, (4)Deictic（指示、ポインティング）に分類している。Ekman が言葉の代りに用いる表象動作を強調しているのに対し、その中の例示的動作に分類される動作に注目し、こちらはマルチモーダルであることを前提にしている。表象動作は典型的なボディランゲージであるため、しばしば画像処理を使ってその解釈の試みがなされるが、他の入力デバイスを人間に置き換えたに過ぎないことがあり、かえって自然さを損なうことがある。

2.2 人間対機械の場合

マルチモーダル・インタフェースの例としては R. Bolt[3] が、発声 (speaking), 動作 (gesture), 視線 (looking) を3つのモードとして挙げてこれらのモードを協調させたり同時に使うことによって、気軽さ (unburdening), 加算による頑強さ (summation), 冗長性 (redundancy) が得られるという利点を説いている。このように複数種類のメッセージを組み合わせて、人間が本来伝えようとしている、あるいは自然に伝わる大事なメッセージを理解しようというのが、マルチモーダル化の基本的な考え方である。

著者らは以前、人間の各部動作からのメッセージを非接触で抽出、統合、解釈してコンピュータ

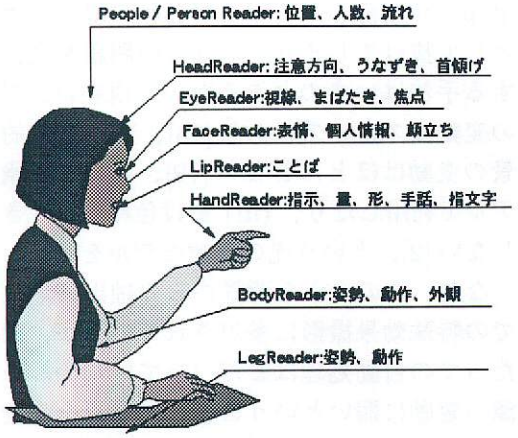

図 1: ヒューマンリーダ：人間各部から発するメッセージ

図 2: 電子秘書とのインタフェース：頭部動作からYes/Noとメニュー選択を区別認識する

と自然なインタラクションを可能とする新しいマンマシンインタフェースのコンセプト、ヒューマン・リーダ (Human Reader) を提案し [9]、それに基づいて数々のサブシステムを開発し、また統合した [10, 11]。すなわち、図1のように各部の動きに注目して身体の部分から発するメッセージを抽出、統合して自然なインタフェースを実現することを試みた。図2は頭部の動きだけを認識解釈するヘッドリーダ [12] を用いて試作した、ビデオメールの到着などを知らせてくれる電子秘書 [10] である。ここでは頭部の動作がもつメッセージのマルチモダリティ（Yes/No の応答と、視方向による興味明示）を処理し簡単な対話とメニューによるシステムの可能性を示した。また、図3は手の動作認識と音声認識を組み合わせたマルチモーダル・インタフェースの試作例、Finger-Pointer[16] である。手形状が示す情報の多義性（指示場所、指文字など）を音声情報により補完して正しく解釈し、プレゼンテーション用インタフェースとして示した。

これらは画像処理を用いることで、非接触、非装着なインタフェースの方向性を示すことができたが、ユーザが着席していることを前提としている。その典型例は会議の場面であり、仮想臨場感通信会議のような応用が考えられる [13]。

最近では画像を使うことの優位性を活かして、ユーザの自由度をひろげた、マルチモーダルなインタフェース環境についての提案も数多くある。例えば、Pentland らのスマートルーム [2]、長尾らの Social Agent[14]、Waters らの DEC-Face のプロジェクトなどは、ユーザが移動することを前提にしたインタフェースを試作している。

3 マルチモーダル・インタフェースの画像処理

インタフェースのために画像処理を利用する際には、次のような特徴を持っていることが望ましい。

1. 実時間性
2. 発展性
3. 頑健性

しかしながら、現在使われているアルゴリズムは次のような特徴をもつ。[2]

1. 単純性：実時間で処理できるように複雑な処理は使われない。センサとしては 50 ～ 100pps 程度の入力が望ましいが、汎用ビデオカメラを用いると最大 30pps しかとれない。画像処理が入るとさらに処理時間がかかり 10-15pps 程度になる。
2. 汎用性：アルゴリズムの変更、ハードウェアの進展に合わせた修正などを可能にするに

[2] これらの特徴が望ましいということではない。現在の研究段階がこの程度であり、これらは製品普及の観点からは阻害要因ともいえる、ということである。

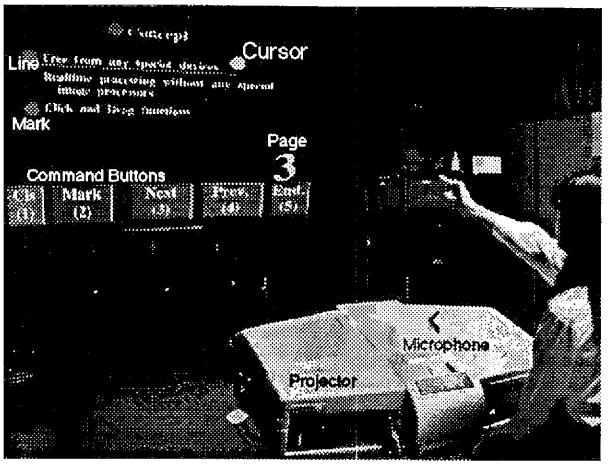

図 3: Finger-Pointer: 手と声を使ったプレゼンテーション・インタフェース

は、処理スピードを犠牲にしてでもなるべく汎用なプログラミング言語を用いる。

3. 制約的：利用者、人数、背景（ブルーバックも含む）、照明条件、動作、対象部分、精度（解像度）、シナリオなどを限定して、処理速度と頑健性を確保している。

以下、具体例を引用しながら各ステップで用いられる画像処理について述べる。

3.1 動作主体の抽出と認識

まず動作主体となる人間やその部分を抽出する必要がある。

3.1.1 主体の背景からの切り出しと追跡

処理対象となる部分を背景から切り出すには、(1) 最初から背景は変化しない、(2) 部分の形状あるいはパターンが既知でモデル化できる、(3) 対象部分の撮影位置は固定である、といういずれかの仮定で処理をしている。

まず、背景が変化しない場合には、最初に覚えた背景パターンを、次々に入力される画像から各画素ごとに引き算して、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さまざまである。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手法はいまのところない。現在は、(i) 背景の変動はスムーズである [15]、(ii) 統計的には背景の変動はほとんどない [12]、という経験的なモデルを利用したり、(iii) 影は色相には大きく影響しない [2]、という光の反射モデルを利用している。なお、このような背景からの抽出は最近の映画での特殊効果撮影に多用されているが、細部にわたっての自動処理はまだできていない。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）を用いるなどの工夫も提案されている [16]。

対象部分の切り出しには対象の形状やパターンのモデルが使われることも多い。例えば、背景差分で対象部分が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。しかしながら実時間で処理するにはモデルとのマッチングを高速に行えるハードウェアを用いるのがほとんどである。なお、対象を人物とせず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。その際、マーカの識別が課題であるが、形状のモデルや連続性から推定できるとされている。

さらに、対象部分の位置がほとんど変化しないという条件をつけて特定位置に対象部分があることを前提とした処理も使われる。表情のモーションキャプチャに使われる、ヘルメットにつけたカメラがその例である [13]。

主体となる対象部分はインタラクションの中で運動しているので、次は追跡する必要がある。運動には慣性など物理的な性質が使えるから、追跡する場合にも次の画像中の位置を大まかに推定することができる。これを使うと先述の切り出しも含めて処理する窓を小さくでき、高速化をはかることができる。

3.1.2 主体位置の特定

主体が画像中で抽出できても、使い方によって必要な情報になっている訳ではない。目的に応じ

て、3次元世界にマッピングするなどの必要がある。実際に行われているのは、(i) カメラ画像座標とほぼ1対1の2次元世界へマッピングした位置、(ii) 3次元世界ではあるが床に接地したような制約条件下の位置、(iii) 空中を自由に動く指先のような完全3次元世界での位置などへの変換がある。(i), (ii) は1台のカメラでも可能であるが、(iii) は原理的には2台以上のカメラを用いて三角測量で3次元位置を決定することになる。いずれも3次元計測の問題になり、注意深いカメラのキャリブレーションが要求されるような応用もある。マウスを使うときのように手元の座標と画面の座標がずれていても問題のないインタフェースもあり (McNeill の Metaphoric に相当する)、その際はフィードバックをうまく設計するとキャリブレーションが大きな問題とならなくなる [16]。

物体位置に加えて、さらに頭の向きや手の方向のような3軸周りの向きの特定が必要となる場合がある。これらは物体上の2点以上の3次元位置から決定するのが基本となる。また、物体の見え方で方向を決定する方法が注目されているが、いかに高速化するかは1つの課題である [20]。

3.1.3 人物の識別

システムの反応動作はユーザごとに変わるのが賢いインタフェースといえよう。ユーザごとにカスタマイズしても、使っている人が誰かわからないと困る。そのための人物の識別処理は重要であり、計算機による顔の認識は長い歴史がある。最近では顔画像を1つのベクトルパターンとみなし、固有顔という特異値分解したパターンで表現する方法が注目され高い認識率を誇っている [20]。これも顔の領域をうまく抽出しているかどうかが全体の性能を左右するので、部分の切り出しは重要である。

3.1.4 人数の識別

いまのところ、人が入れ替わったり頻繁に移動するような、複数の人との積極的なインタラクションのためのインタフェースを構築するための処理手法は見あたらない。しかし、例えば部屋にいる人数を知っていることはシステムの動作決定に役立つ。モニタリングのための人数の計数処理が報告されている [10]。

3.2 動作の認識と解釈

動作の認識と解釈は画像処理というよりパターン認識の範疇である。動画像列モデルと直接マッチングすることもありうるが、計算量が膨大となり現実的でない。一旦は動きパターンに変換してそれを識別するべきである。

前述の電子秘書の例では、頭部の回転角の変化をチェックして1点を凝視しているか、頭を振っているかを区別した。動作に関わる部分の関節角の相互の関係を使った位相によるパターンの識別などがあるが、いま注目すべきは音声認識で用いられている HMM(Hidden-Markov-Model) を使って動きの解釈をする手法である [17]。運動パターンや手話の認識に用いられている。

4 MIC インタフェース

なぜ私たちは人間同士で対話しているのと同じような“自然なインタフェース”を欲しがるのだろうか。次のような点を期待しているのではないだろうか。

1. 自然なインタフェースは学習が容易あるいは無用。コンピュータや機械の使い方を学習することはある程度必要である。しかし、人間に適した使い方でないと忘れやすかったり、学習の進展がにぶい。
2. 自然なインタフェースは人間の処理速度に合わせた操作が可能。コンピュータや機械の（情報）処理能力は高速かつ大量であり、人間の処理能力や状況に合せたインタラクションが必要である。
3. 自然なインタフェースをもつ機械を使うと創造性が刺激される。コンピュータや機械を道具として使うときに道具を愛でることでその能力を引き延ばすことができる。

これらを考えるとき、今後はマルチモダリティ (M) だけでなく、賢いインタラクティビティ (I) とクリエイティビティ (C) をも兼ね備えた MIC インタフェースの研究が必要ではないだろうか？これはインタフェースとタクスプログラムが混然となってくることになるかもしれない。

5 あとがき

モーバイルコンピューティングが日常的になると思われる将来は、動作の入力には画像による非接触センシングに固執せず装着型の入力デバイスを活用したマルチモーダルインタフェースが多用されることになるだろう。もちろん腕時計くらいの大きさの装置でなくては困るが。また、動作だけでなく強度、触覚あるいは応力を伝えることができる手軽な入力デバイスなども開発されるだろう。しかしそれでもなお、線でつながっていない他人の動作や環境のセンシングは状況理解のために必要であり、画像処理の活躍する場面は決してなくならないと考えられる。

最後に、インタフェースは非常に重要な研究課題であるが、最初に述べたように何のためのインタフェースかを忘れると研究目的が主客逆転し、迷走におちいる恐れがある。ATRの我々の研究室では、人間同士のコミュニケーションにおいて相互理解、創造的発想、知識共有を支援するインタフェース・エージェントの研究を開始した。このエージェントは、例えばミュージアムへ訪問したときのように膨大な知識ベースを前にしたコミュニケーションに介在してわたしたちの創造的な活動を刺激する環境を提供してくれるだろう[18, 19]。現在このエージェントにMICインタフェースを装備させるべく研究をすすめている。

謝辞

日頃ご指導いただくATR知能映像通信研究所 葉原耕平会長ならびに中津良平社長、また議論いただく研究所各位に感謝します。ヒューマンリーダプロジェクトはNTTの末永康仁氏ほか多くの方の協力でNTTヒューマンインタフェース研究所にて行なわれた。

参考文献

[1] R. Nakatsu and K. Mase: "Tutorial Course on Life-like Believable Communication Agents", SIGGRAPH96, New Orleans(1996).

[2] A. P. Pentland: "Smart rooms", Scientific American, pp. 54–62(1996).

[3] R. A. Bolt: "The integrated multi-modal interface", 信学論, **J70-D**, 11, pp. 2017–2025(1987).

[4] 鳥脇純一郎: "ヒューマンインタフェースと画像処理", コロナ社(1995).

[5] 末永康仁: "人物像を読む－ヒューマンインタフェースのためのコンピュータビジョン", 信学誌, **78**, 8, pp. 800–804(1995).

[6] P. Ekman: "Biological and cultural contributions to body and facial movement", The Anthropology of the Body (Ed. by J. Blacking), Academic Press Inc., New York, pp. 39–84(1977).

[7] 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).

[8] J. Cassell et. al.: "Animated conversation: Rule-based generation of facial expression, gesture & spoken intonation for multiple conversational agents", SIGGRAPH94 Proceedings, Orlando, FL, pp. 413–420(1994).

[9] K. Mase, Y. Suenaga and T. Akimoto: "Head reader: A head motion understanding system for better man-machine interaction", IEEE proc. SMC, pp. 970–974(1987).

[10] 間瀬健二: "動画像処理を用いた新しいマンマシンインタフェースの研究", PhD thesis, 名古屋大学学位論文(1992).

[11] 末永康仁, 間瀬健二, 福本雅朗, 渡部保日児: "Human reader: 人物像と音声による知的インタフェース", 信学論, **J75-D-II**, 2, pp. 190–202(1992).

[12] 間瀬健二, 渡部保日児, 末永康仁: "ヘッドリーダ: 画像による頭部動作の実時間検出", 信学論, **J74-D-II**, 3, pp. 398–406(1991).

[13] K. Fumio, T. Miyasato and T. Terashima: "Virtual space teleconferencing", in Proc. IEEE Int'l Workshop on Robot and Human Communication, pp. 205–210(1995).

[14] K. Nagao and A. Takeuchi: "Social interaction: Multimodal conversation with social agents", In Proc. AAAI-94, **1**, pp. 22–28(1994).

[15] A. Sato, K. Mase, A. Tomono and K. Ishii: "Pedestrian counting system robust against illumination changes", SPIE, VCIP'93, **2094**, pp. 1259–1270(1993).

[16] M. Fukumoto, K. Mase and Y. Suenaga: "Finger-pointer: Pointing interface by image processing", Comput. & Graphics., **18**, 5, pp. 633–642(1994).

[17] 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", 信学論(D-II), **J77-D-II**, 7, pp. 1311–1318(1994).

[18] Y. Sumi, K. Mase and K. Nishimoto: "Facilitating human communications in personalized information spaces", AAAI-96 Workshop on Internet-based Information Systems(1996).

[19] 門林理恵子, 間瀬健二: "新しいコミュニケーション環境としてのMetaMuseum", マルチメディア通信と分散処理ワークショップ論文集, 情処, pp. 71–78(1995).

[20] A. Pentland, B. Moghaddam, and T. Starner: "View-based and modular eigenfaces for face recognition," CVPR'94, pp.84–91 (1994).