原稿No.GA03-01

シリーズ：画像処理技術の基本を学ぶ（第6回）

ヒューマンインタフェースのための画像処理

インタフェースにおけるインタラクションのデザイン

㈱ATR知能映像通信研究所

間瀬　健二

1997年1月12日

冒頭から脱線して申しわけないが、上記の日付で何を思い出すだろうか。雑誌でも特集が組まれた[1]りしたので記憶に新しい方もいるだろう。アーサーC. クラークのSF小説「2001年宇宙の旅」に出て来るコンピュータHAL9000の誕生日である※1。本誌で前回「ヒューマンインタフェースと画像処理」という記事を書いた[2]とき、ある意味では理想的なヒューマンインタフェースを備えたHALを引き合いに出したことは記憶に新しい。そしてまた、ちょうどこの記事の締め切り前にHALの誕生日をむかえた。当所では映画「2001年...」をみんなで見ようと、当研究室の実験用の170インチのスクリーンの前に有志（物好きともいうらしい）30名くらいが集まってパーティとなった光景が、タイプする手にオーバーラップしてくる。

だれも「2001年...」を予言と考えるコンピュータ科学者はいないだろう。しかし、HALのように対話できるコンピュータを作り育てることを目指している多くの研究者のうち誰かが、もしかしたらこの瞬間、大事なプログラムを作っているかもしれないと想像をめぐらすのは楽しい。それはどんなアーキテクチャ上で動作し学習をしていくのだろう？またどんなセンサーやアクチュエータを備えているのだろうか？そして、自分の研究もそのような未来に道を作るような仕事であって欲しいと願うのは私だけではないだろう[3]。

※1 映画版では、5年早く1992年1月12日に設定されている

コンピュータとの対話シーンの変化

さて、近年のコンピュータの進歩は、クラークとキューブリックが描いた未来のコンピュータとはいろいろな面で違う未来の様相を示しつつある。いまやコンピュータと対話する姿として、写真1[4]のように机に座ってキーボードを叩いたり、モニタディスプレイに向って音声で会話しているのさえ、未来の光景としては物足りない。小さなバッジのようなコンピュータがいつも一緒にいて執事のように世話をやいてくれたり、あるいはコンピュータが提供する空間の中に自分自身を没入させてコンピュータを使う光景が目にうかぶ。

本文では、そのようなコンピュータとのコミュニケーション形態において画像処理がヒューマンインタフェースに果たす役割を、実例を示しながら述べる。画像処理技術そのものではなく、画像処理をヒューマンインタフェースに応用する際の課題とその対処について事例を用いて述べることにする。

写真1　デスクトップコンピュータとの対話シーンの一例：電子秘書との対話を頭部の動作で行う。Yes／Noとメニュー選択を区別認識することができた。コンピュータと対話するのに、椅子に座らないといけないのだろうか？

マルチモーダルインタフェースと画像処理

インタフェースにおける画像処理の役割を考えるときに、まずコンピュータにとっての目の機能による情報取得があり、次に人間の目の機能にとっての情報提示がある。これらはコンピュータシステムへの情報入力とシステムからの情報出力にあたるが、入出力それぞれにマルチモーダルな情報のやりとりを考えることで、操作性や認知的負担度などを軽減することができる。

第1表はマルチモーダルインタフェースの特徴を表にまとめたものである[5]。入力にはユーザのアクションに関する冗長性が許されたり入力信号に対する

第1表　マルチモーダル・ヒューマンインタフェースの特徴

システムへの入力	
冗長性	複数メディアを使うことで、ユーザは同じことをさまざまに表現することが可能である
気軽さ	ユーザは普段慣れているメディアを使うことができる
頑健性	複数のメディアからの情報を統合するので、各メディアからの情報は完全でなくてもいい
システムからの出力	
論理性	文字や表を中心とする論理的な表現が可能である
直感性	図やグラフを用いることで数値的な比較を直感的に表現できる
実感性	写真、動画によって実物に近いものを実感できる
情感性	視覚的、聴覚的なメディアを利用することで情感的な表現が可能になる
対話性	情報を複数メディアで表現できるので、状況に応じた情報の利用や体験により正確な伝達が期待できる

第2表　ジェスチャが生じるメッセージ

ジェスチャ・メディア	メッセージ／機能	動作例
ロケータ	空間中の場所指定	指示、頭の向き
セレクタ	リストからの選択	指文字
バリュエータ	量や大きさの指示	ダイアル操作
イメージャ	イメージや概念の提示	ボディランゲージ、表情

システムの頑健性が生まれ、そして、出力にはユーザの理解を高める直感性や感性に訴える情感性を期待することができる。写真2はそれを実現した1例で、音声と手のジェスチャを組み合わせたマルチモーダルインタフェースの例である[6]。手の先を2台のカメラで撮影し、画像処理で指文字形状や指示方向の情報を得ている。また、動作に同期して音声認識モジュールが「あれ」、「ここまで」、「これだけ進めて」という単語を認識している。第2表に示すように、手はその使い方でいろいろな意味をもつマルチモーダルなメッセージ発生源であるから、音声と組み合わせることによって動作解釈における頑健性をもたらすことができる。

このように、入出力それぞれについてインタフェースをマルチモーダル化することによる利点は大きい。そして、さらに入力と出力の組み合わせ、すなわちこれらの相互作用（あるいはインタラクション）の過程にもインタフェースを改善する効果を期待することができる。第1表にあるように入力においては状況依存性を利用することができ、出力においては対話性をもたらす環境を作り出すことができると考えられるからである。

インタラクションのデザイン

例えばコンピュータゲームにおける経路選択やスクリーン上のメニュー選択において、ユーザには必ずしも多くの自由度が与えられているわけではない。しかし、あるストーリのなかでユーザ自身が知らないうちに限定された世界に誘導されていけば、そのなかでの選択枝や自由度の制約を不満に思うことは少なくなるだろう。

人間はマルチモーダルな情報発信源であり、どのモダリティで情報を発しているかは状況に大きく依存する。したがって画像処理でジェスチャ認識をしてそれに意味付けをして解釈する場合に、まず状況を認識しておくことは有効である。コンピュータがその状況を作り出すか、あらかじめシナリオをしっていれば、画像処理で認識すべき対象を簡単化したりすることができる。

状況によってジェスチャの解釈が異なる場合に、我々はそのインタフェースをいかにデザインすればよいだろうか？以下では、ジェスチャ認識用画像処理モジュールPfinderを用いて、ATRで作成した全身ジェスチャインタフェースの2つのシステム例を紹介する。

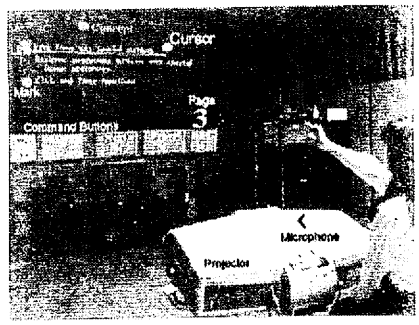


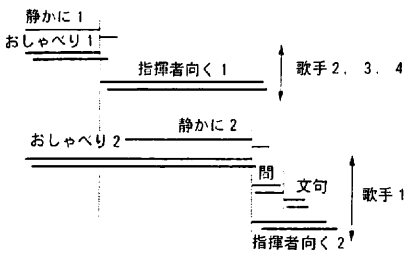
写真2　Finger-Pointer：手と声を使ったプレゼンテーション・インタフェース

"Pfinder" はMITメディアラボのAlex Pentland教授らのグループで開発された単眼画像のリアルタイム処理によるジェスチャ認識プログラムであり[7]、サイバー犬と戯れる仮想環境を提供するALIVEシステム[8]などに用いられている。

ここではPfinderを一つの入力デバイスとしてその入力信号をどのようなコマンドに割り当てるかを中心に議論する。Pfinderは基本機能として、ユーザの床面上での位置、ユーザの屈伸、手の上げ下ろしなどの状態を認識することができる。画像処理の詳細については[7, 8]などを参照されたい。

SingSongにおけるインタラクション

"SingSong" は人間の俳優が指揮者を演じ、4人のコンピュータ俳優（CG）が歌手を演じて曲を演奏するというシナリオによる臨場感あるインタラクティブシステムである。指揮者のジェスチャをもとにシナリオが進行する。このような、俳優同士のかけあいは相互の動作のタイミングを記述しなければならない。しかし、一列に並べ



第1図　SingSongの1シーンを記述するInterval Scriptsの図式表現

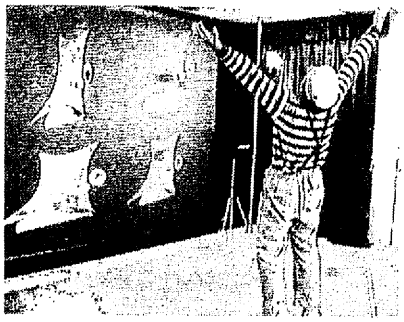
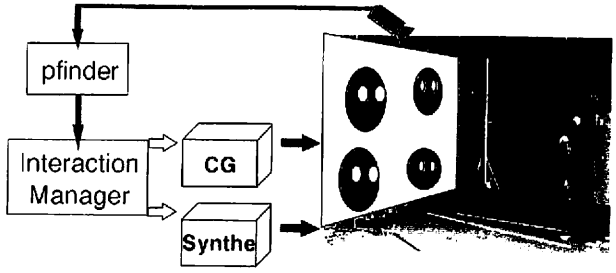


写真3　SingSongの1シーン：人間俳優がCG俳優の歌手を指揮している

られたイベントのタイミング制御だけではなんのおもしろさもないので、動作やメニュー選択に自由度を持たせるのであるが、すると途端にイベントの展開をコンピュータプログラムで記述することが困難となる。

そこで、我々はAllenのTemporal Interval Logic(時区間論理)を利用したIntervalScriptsというインタラクション記述言語を開発した[9]。ジェスチャ認識プログラムをセンサとしCGの動作表現をアクチュエータとするようなインタラクションは、イベントの開始や終了のタイミングが不確定である。Interval Scriptsは2つのイベント相互の時間的関係の集合でインタラクション全体を記述することを目的としている。

第1図はその記述の一例である。「コンピュータ俳優はおしゃべりをしている間、指揮者に注目し、指揮者の手があがったら、指揮者のほうをむく…」という流れのインタラクションを、時区間(Temporal Interval)を線分で標記しながら図式に表現している。このようにすることで、「Aの直後にBを起動する」とか「Cの間、Dをする」という記述が容易となった。



第2図　実験システムの構成

VisTA-Walkにおけるインタラクション

VisTA-Walk[10]は弥生時代の遺跡集落を再現した仮想空間のウォークスルーシステムで、集落の変遷を視覚的に擬似体験して理解や知識発見を支援することを目的としている。疑似体験をする際に、自然な歩行動作に近い動作で、臨場感あふれる空間を歩き回れる環境を提供するためにPfinderを用いてインタラクションを設計している。

写真4がその様子であるが、ユーザが立つ中立位置から足を踏み出すと空間を移動することができる。スクリーン前の空間が有限なため、どんどん歩いていくということはできないが、Pfinderが出力する立つ位置をジョイスティックのように解釈して空間を移動できる。実際にやってみると歩いているような雰囲気を味わうことができるから不思議である。

VisTA-Walkはそのほかに手の指示動作や振り上げなどで空間中のオブジェクト(この場合、住居など)を選択し、それに関する情報を閲覧できる。SingSongとは異なり、これらのジェスチャによるインタラクション処理には通常のイベント手続きでプログラミングを

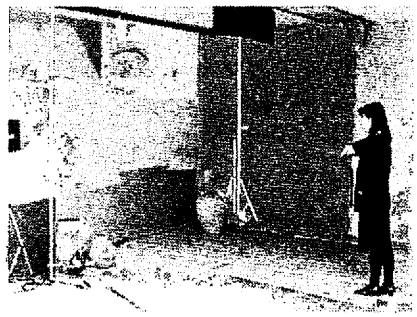


写真4　VisTA-Walk：仮想弥生集落のウォークスルー

している。その理由は、現在のところInterval Scriptsにジョイスティックのようなバリュエータセンサが定義できていないからである。つまりSingSongではPfinderをロケータおよびセレクタセンサとしてデザインしてある。これについては今後の課題である。

SingSongおよびVisTA-Walkのシステム構成は第2図のようになっている。170インチ・スクリーンの上部に設置されたカメラの映像を画像処理しインタラクションマネージャを介して画像生成部を制御して表示している。画像処理にはIndy R4400、全体の制御と画像生成にはOnyx R10000-IRを用いている。

あとがき

ヒューマンインタフェースに画像処理を利用する際に問題となる、インタラクションのデザインを中心に解説した。画像メディアを用いたマルチモーダルインタフェースはユーザに自然さ

を提供するが、画像処理部の困難さや外乱や雑音に弱いという問題がしばしば指摘される。確かに、現在もPfinderは照明の変化、背景の変化、2人以上ユーザなどに対応できないという問題をもっている。しかし、インタラクションのデザインを注意深くして、状況にあわせて画像センサの出力を解釈するようにすればもっと頑健なシステムができるのではないかと考える。また、システム設計上からは、画像センサにこだわらず他種センサとの組み合わせを考えることも必要である。

参考文献

1) S. Garfinkel, "Happy Birthday, HAL", WIRED, 5.01,, Jan. 1997.
2) 間瀬, "ヒューマンインタフェースと画像処理", 画像ラボ, pp. 34-37, 日本工業出版, Jan. 1991.
3) D.G. Stork, editor, "HAL's Legacy", MIT Press, 1997.
4) 間瀬, "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", 名古屋大学学位論文, Mar. 1992.
5) 宮里, 間瀬, "ネットワーク利用者を支援するマルチモーダルヒューマンインタフェース", 情報処理, 38, 1, pp.42-47, Jan. 1997.
6) M.Fukumoto,K.Mase,and Y.Suenaga, "Fingerpointer : Pointing interface by image processing", Comput. ε Graphics., 18, 5, pp. 633-642, May 1994.
7) C. Wren, A. Azarbayejani, T. Darrell, and A.P. Pentland, "Pfinder:Real-Time Tracking of the Human Body", In 2nd International Conf. on Automatic Face and Gesture Recognition, Killington, Vermont, Oct. 1996.
8) A. P. Pentland, "Smart rooms", Scientific American, pp. 54-62, April 1996.
9) C.S. Pinhanez, K. Mase, and A. Bobick, "Interval Scripts: a Design Pradigm for Story-Based Interactive Systems", In ACM-CHI'97, pp-287-294, Georgia, Mar. 1997.
10) 門林, ネータル, 間瀬, "身振りインタフェースを用いた集落変遷シミュレーションシステム", 1997年信学会総合大会基礎境界, p-396, Mar. 1997

【筆者紹介】

間瀬健二
(愛知県出身)
㈱ATR知能映像通信研究所
第2研究室　室長
〒619-02 京都府相楽郡精華町光台2-2
TEL : 0774-95-1440
FAX : 0774-95-1408

人物紹介シリーズ

この項目はこの本文とは関係なく本誌に登場した執筆者・エンジニヤを紹介するシリーズです。

見持圭一 (昭和32年3月19日生・兵庫県出身)
三菱重工業㈱　高砂研究所　電子技術研究室
〒676　兵庫県高砂市荒井町新浜2-1-1
TEL:0794(45)6809　FAX:0794(45)6088

＜趣味＞　ゴルフ、読書
＜定期購読誌・紙＞　朝日新聞
＜家族構成＞　妻、長女、二女
＜主なる業務歴および資格＞
昭和56年三菱重工業株式会社入社。高砂研究所、基盤技術研究所（横浜市）を経て、平成7年4月より再び高砂研究所勤務。画像計測・認識、動画像処理、ロボットビジョン（特に3次元視覚情報処理技術）の研究開発に従事。

E-mail : kemmotsu@wt.trdc.tksg.mhi.co.jp

＜執筆タイトル＞　最適センサ配置の設計

三菱重工業株式会社
＜代表者＞　増田信行
＜本社住所＞　〒100　東京都千代田区丸ノ内2-5-1
TEL:03(3212)3111
FAX:03(3212)9800
＜資本金＞　2647億円
＜年商＞　2兆5030億円
＜従業員数＞　4万1957名
＜事業内容及び会社近況＞
船舶・海洋、鉄構建設、原動機、原子力、機械、産業機械、航空機・特車、汎用機、冷熱の9事業本部によって構成され、700の製品を有する機械のデパートです。詳しくはホームページhttp://www.mhi.co.jpをご覧下さい。

寺田賢治 (昭和42年5月25日生・京都府出身) 徳島大学　工学部　知能情報工学科　講師
〒770　徳島県徳島市南常三島町2-1　TEL:0886(56)7499　FAX:0886(56)7319

＜主なる業務歴および資格＞
平成2年3月慶應義塾大学理工学部電気工学科卒業。
平成7年3月同大大学院後期博士課程修了。同年徳島大学工学部助手、現在、同工学部講師。
平成4年から平成7年まで日本学術振興会特別研究員。画像計測に関する研究に従事。工学博士。
電子情報通信学会、計測自動制御学会等の会員。
＜執筆タイトル＞ 3次元センサーを用いた人の顔の自動識別システム