ヒューマンインタフェース　95－23
コンピュータビジョンと
イメージメディア　129－23
（２００１．９．１４）

マルチモーダル対話における視覚の役割とその応用

長谷川修

産業技術総合研究所 脳神経情報研究部門

o.hasegawa@aist.go.jp

あらまし　近年、高度化・複雑化の一途を辿る機械や情報機器類と人間との乖離が進み、社会問題化しつつある。そこで現在、これまで以上にわかりやすく、親しみやすいマンマシン・インタフェースを求める声が強まっており、その候補の一つとして、人の五感や身体を複合的に利用するマルチモーダル対話インタフェースへの期待が高まっている。中でも、機械／情報システムに視覚や聴覚、表情やジェスチャを表出する「身体」を与え、人間にとって自然な対話形式でシステムとのインタラクションを図るアプローチは、最も野心的かつ有望な方向性の一つとして国の内外で活発に研究されている。本稿では、マルチモーダル対話における「視覚」の役割について検討を加え、また筆者らの最近の研究の概要を述べる。

Roles of Visual Functions in Multimodal Interface and Their Applications

Osamu Hasegawa

Neuroscience Research Institute,
Advanced Industrial Science and Technology

Abstract　Recently, highly developed equipments cause severe divides / separations between human users and machines. Thus researchers are strongly required to design and develop an easy to understand man-machine interface as soon as possible. A “Miltimodal Interface” with visual and audio recognition functions and a body for facial and gesture expressions is considered to be one of the most promising approaches to solve the problem. In this paper, we consider roles of the computer vision in multi-modal interfaces and describe outline of our work.

1　はじめに

情報通信技術の進歩やネットワーク・インフラの整備により、私たちをとりまく情報環境は近年急速に発展しつつあり、今後そのスピードはさらに加速されると思われる。その一方で、高度化・複雑化の一途を辿る機械や情報機器類と人間の乖離も進んでおり、今日ではそれが深刻な社会問題となりつつある。

そこで現在、これまで以上にわかりやすく、親しみやすいマンマシン・インタフェースを求める声が強まっており、その候補の一つとして、人の五感や身体を複合的に利用するマルチモーダル・インタフェースへの期待が高まっている。中でも、機械／情報システムに視覚や聴覚、表情やジェスチャを表出する「身体」を与え、人間にとって自然な対話形式でシステムとのインタラクションを図るアプローチは、最も野心的かつ有望な方向性の一つとして国の内外で活発に研究されている [1, 2, 3, 4, 5, 6, 7]。

さらに最近では、そうしたシステムが持つべき「知能」や「身体」の研究が進展し、それが脳科学、認知科学、発達心理学、計算機科学、数理科学、ロボティクスなどと発展的に融合されて「人を知り、その知見に基づいて人と親和性高く共生する知的な存在を創る」ことを目標に掲げた大きな研究の流れが形成されつつある [9, 10, 11, 12, 13, 14, 15]。

本稿では、以上のような状況を踏まえつつ、マルチモーダル対話における「視覚」の役割について検討を加え、また筆者らの最近の研究の概要を述べる。

2　人は何を見ているか

まず、人どうしが対面対話（マルチモーダル対話）する際の、視覚の役割に関する知見を概観する。なお一般に成人が対話する際の視覚の振舞いは、相手との関係、経験や興味、背景にある文化などに大きく依存し、そこから基本的・本質的な知見を見い出すことは難しい。そこで、そうした影響の少ない乳児や幼児に着目し、彼らが何に目を向け、そこから何を獲得し、それを他者とのコミュニケーションにどのように生かしているのかを検討することとする。

乳児が「顔」に選択的に顔を向け、反応することは良く知られているが、これはそうした能力が遺伝子レベルに刻み込まれた先天的なものであることを意味している。乳児のこうした行動は、養育者（親）や周囲の人に養育行動を起こさせるだけでなく、自らの知的発達のため、人（顔）に対して優先的に注意を向けるようプログラムされていると考えることができる。

また、乳児は他人の表情を真似ることができる一方で、脳の神経解剖学的研究から人の感情（情動）と表情表出の間には一定の関係があることが知られており、これらから、他人の表情を自らの表情と比較することにより、「他人の感情を察する」能力を先天的に備えているとも考えられる[1]。以上のような乳児の先天的な能力は、乳児が社会的な存在として発達するための基本的かつ重要な「方向づけ」を与えていると考えられる[2]。

乳児期から幼児期にかけては、顔の学習／識別の能力が発達すると同時に、様々な文脈における表情や視線の意味の理解と学習が進むと考えられる。言語の獲得が始まるのもこの頃であり、この時期は、他者とのスムーズなコミュニケーションや、高度な社会性の獲得のために必要なスキルの形成期と言えよう。成人の豊かな知性や運動能力は、こうしたスキルを基盤に、時間をかけて徐々に獲得されると思われる。

3　機械は何を観るべきか

次に、インタフェース構築の観点から視覚に求められる役割について考察する。

一般にインタフェース研究の目標は「機械を使う人間の側に立った人間中心のインタフェースをいかに実現するか」にあると言え、これまでに、人間の感覚や直観に訴えて「少ない予備知識で操作（コマンド入力）が可能な」インタフェースが数多く提案され、成果があげられてきた。

しかし近年、機械や情報システムの多機能化・高機能化が進み、直観的な表現やそれらの組合せでは理解や操作が困難な機能も多く見られるに至っており、現在、

- 直観的に理解しやすい表現（コマンド）を用意して「入力を待つ」アプローチから
- システムが状況を判断し、相応しい機能やサービスを能動的に「提供する」アプローチへ

とパラダイムの転換が図られている。

冒頭に述べた対話型のマルチモーダル・インタフェースは、まさにそうした方向性を目指すものの一つと言え、その中でも「視覚」は利用者の状態や振舞いに関する情報を能動的に得ることのできる手段として重要な役割を果たすことが期待されている。

以下では、そうした観点から「顔を観る」、「ジェスチャを観る」の二点について、「視覚」の果たすべき役割や今後の課題について考察する。

[1] こうした能力は、集団社会において無用な争いを避けるために有用であり、進化の過程で獲得されたのではないだろうか。

[2] 脳の機能的な障害により、この能力を持たない方を自閉症と呼ぶ。自閉症の方は他者を模倣したり共感することが苦手で、他者とのコミュニケーションに困難を伴い、社会性の発達が遅れると言われている。なお自閉症と知的障害は別のもので、知的レベルの高い自閉症の方も多く存在する [8]。

3.1 顔を観る

インタフェースへの応用を前提とした顔画像の処理には、＜顔の抽出、人物の識別、年齢の推定、性別の識別、顔（頭）の向きの推定＞、＜顔（頭）の動きの認識、表情の認識、視線の認識＞などがあり [21]、これらは括弧で分けたように、大きく「計測の問題として扱えるテーマ」と「計測の後に認識が期待されるテーマ」とに分けられる。

これまでのところ前者に比べて後者の進展が遅れているが、これは後者の最終目標が表情や視線、頭の動きからの相手の「心情や意図の推定」にあり、その表出のされ方が個人差や文化が大きく影響されるため、問題がより一層難しいことによる。

しかし一般に、人の交わす会話は相手の心情や意図を察しつつ行なわれており、その点において、計算機間の通信と一線を画している。いわゆる「機械的」でない、人と親和性の高い対話を人と機械の間に導入するためには、上記の後者の研究の進展はが必須である。

後者の目標の達成には、先に述べた乳児の先天的能力の研究や、乳児から幼児に発達する過程の脳内メカニズムに関する知見が参考になると思われる。つまりそれらはシステム構築の際、「事前に組み込める（おくべき）機能は何か、組み込めない機能はいつどのようにして獲得されるのか、他のモダリティとの関連は」といった点を考える上で、多くのヒントを与えてくれると考えられる。

他にも、最近の＜大脳辺縁系、視床：先天的機能＞、＜前頭葉、側頭葉：後天的学習に関与＞などの研究や、それらの連携による判断や記憶の形成（学習）に関するシステム論的研究などは、かなり示唆的と思われる [16, 17, 18, 19, 13]。

3.2 ジェスチャを観る

ジェスチャ認識の研究も大きく二つに分けられると思われる。一つは「意識的に表出される記号」としてのジェスチャの計測／認識の研究であり、もう一つは「無意識に表出されるモダリティの一つ」としてのジェスチャの認識の研究である。

一般に、ジェスチャは三次元空間内の時系列情報であるため、単眼や二眼では本質的に計測が難しい。そこで、これまでのジェスチャ認識の研究の殆んどが、相手に見せることを意識して表出された記号的ジェスチャの認識となっている。しかし、より自然な対話の実現の観点からは、無意識的に表出されるジェスチャの認識が不可欠であり、今後の研究の進展が期待される。

近年、サルの電気生理学的研究によって運動前野で発見されたミラーニューロン [20] は、そうした研究に多大なヒントを与えてくれるかも知れない。ミラーニューロンは、自らのある「行為」に反応する他、同じ行為を他者が行なった場合も反応するという興味深い性質を持っている。すなわち、自己の行為（ジェスチャ）と他者の行為のマッチングを行なっていると考えられるほか、他者の心の推定にも関与しているのではないかとも言われている。今後もミラーニューロンに関する生理学的研究や、ミラーニューロンとその周辺に関するシステム論的脳研究が注目される。

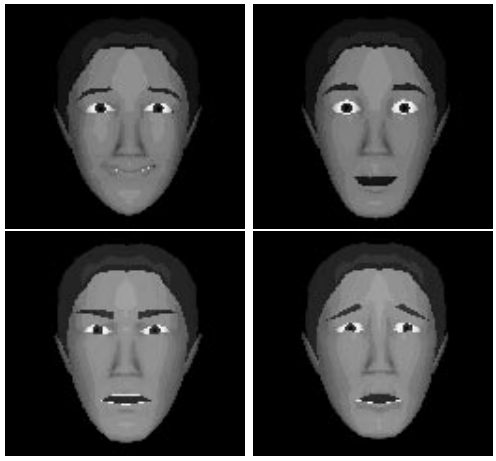

図１：表情の合成例：喜び、驚き、悲しみ（困惑）

図２：３次元位置の表現：（ロボットの身体性（視線や指さしジェスチャ）を活用し３次元空間中の位置を直観的に分かり易く表現することができる。）

4 筆者らの取り組み

以下では、現在筆者らが進めている、人間型ソフトウエアロボット [1, 2, 12] の研究とその応用システムについて概説する。

4.1 人間型ソフトウエアロボット

4.1.1 研究の経緯

人間との共存を目指すこれからの機械情報システムには、実環境の物理的な側面と、人間の作る社会的な

側面の双方において、学習し成長する能力が求められると考えられる。そこで現在筆者らは、人に似た姿をしたCG像に視覚／聴覚／顔表情／ジェスチャ／発話といった人の日常的なモダリティを与え、できるだけ人間に近い形式で、子供が成長するように学習するシステムの研究・開発を進めており、これを人間型ソフトウエアロボット（以下ロボット）と呼んでいる。

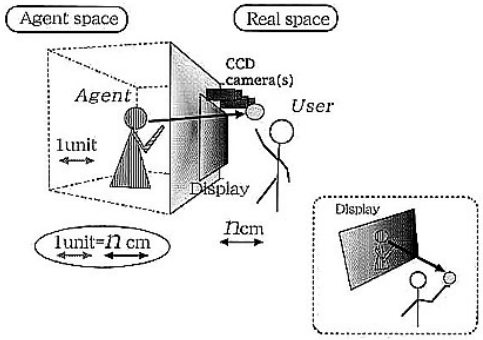

図３：実空間と仮想ロボット空間の融合

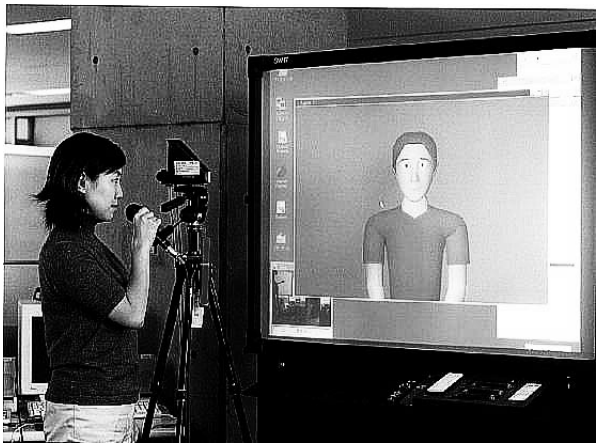

図４：人間型ソフトウエアロボットのプロトタイプと人間の対話の様子

4.1.2　プロトタイプシステムの概要

試作中のロボットは、PC上に画像と音声の認識・合成機能等を統合的に実装し、これに小型ステレオビジョン、マイクなどを加えた構成となっている。

現在の研究のポイントは、実環境から入力される視覚と聴覚情報を統合的学習に置いており、そのメカニズムとして近年研究が急速に進んでいる脳の全頭葉、側頭葉、大脳辺縁系の連携による学習モデルの導入を図っている（詳細については稿を改めて報告する）。

ロボットの基本性能としては、図１に示す顔表情と図２に示す微妙な視線、指さしジェスチャなどを組み合わせて表出可能としている[23]。指さしジェスチャでは、ロボットの描画座標系を実空間に連続的に接続するようにしているため、ロボットは実環境の特定の対象を指し示すことができる。またロボットに「これ／それ／あれ／を見て下さい」という発話や視線とともに指さしをさせると、人間の注意を誘導できることがわかっており、これは人間とロボットどちらからでも共同注意[24]の形成が可能であることを意味している[2]（図３）。

共同注意が成立した状態とは、ロボットと人間の間で実環境の情報を共有した状態であり、これはロボットの実環境に関する情報の学習のために重要である。現在進めている研究は、共同注意を成立させた上で、視覚情報に人間が口頭（音声）でラベルを与え、それをロボットに学習（インターモーダル学習[25]）させるものである[22]。つまり、養育者と乳児／幼児の関係と同じように、人間がロボットに実環境中の様々な対象を提示すると、ロボットは対話を通じてその知識や概念を獲得するといったメカニズムの実装を目指している。図４に、ロボットと人間との対話中の様子を示す。なお現在ロボットは計算機のモニタ上に表示しているが、今後画像表示技術が進めば、任意の位置／場所に表示することが可能になろう。

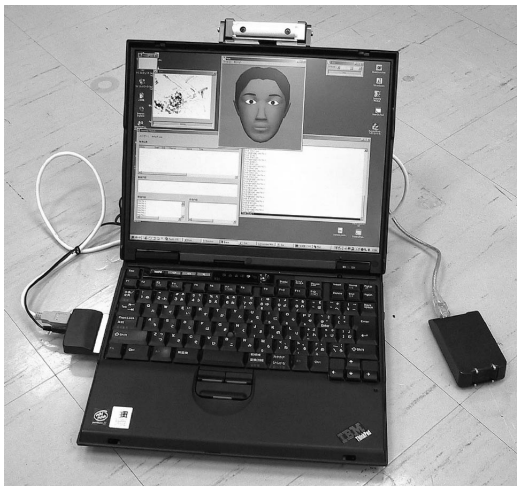

図５：マルチモーダル携帯エージェントの試作機（試作２号機）

図６：小型スレテオカメラ（携帯エージェントシステムのディスプレイ上部に取付けて利用している。CPUにPentiumIII,850MHzを使った場合、320x240ピクセルの

ステレオ画像から毎秒５～６回の距離画像の算出が可能。）

図７：小型ステレオカメラによる画像処理例（上段：左右のカメラからの入力映像。下段左：算出された距離画像。下段右：背景から切り出された人物画像。下段右の画像から人物の識別などを行なう。）

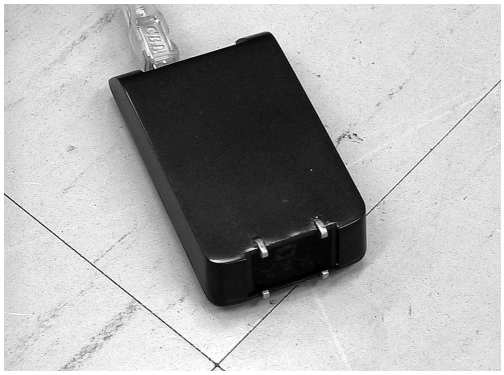

図８：通信デバイス（汎用リモコン。エージェントと家電などとの通信に利用している。制御対象の識別信号を受け、それに対応した制御信号を発信する。）

図９：マルチモーダル携帯エージェントの処理画面例

4.2　応用システム

本節では、人間型ソフトウエアロボットの応用システムについて概要を述べる。

高度情報化社会においては、情報技術を使いこなす者には多くの恩恵がもたらされる一方で、使わない／使えない者には社会的な不平等／不利益がもたらされることが懸念されている。

そこで筆者らは、これまでに構築してきたマルチモーダル対話エージェントを携帯可能な小型端末（ノートPC）に移植し、エージェントとの簡単な対話を通じてネットワーク上の各種資源にアクセスしたり、自宅やオフィスなどに設置した「ホスト」にアクセスして種々の情報を管理・操作することが可能なシステムを試作した。エージェントは基本的にホスト上で稼働しており、外出時に携帯システムに読み込む構成となっている。負荷のかかる計算や、大規模なデータの管理などはホストに処理させ、必要な情報のみ携帯システムに呼び出すといったことも可能である。

この携帯システムは小型ステレオカメラ（図６）を搭載しており、カメラの前に現れた人や物を距離情報を手がかりに切り出して、背景に依存せずにそれらを認識することができる（図７）。これにより認識率が格段に向上した。

現在、利用者の顔の向きを推定し、システムに向かって話しかけられた場合にのみ、エージェントが反応するといったメカニズムの検討なども進めている。

また本システムには家電などとの通信機能を持たせており、これを利用して、例えば初めて見る家電や機器があったなら、まずエージェントにアクセスさせてその機種や機能を調べさせ、ユーザに代わってそれらを操作させるといった機能を実験的に搭載している（図８）。今後、家電や機器のネットワーク化が進み、現在人間向けに書かれているマニュアルの情報がエージェントにも提供されるようになれば、ユーザはメーカや機種ごとに異なるマニュアルを読み、操作法を習得する負担から、かなり解放される可能性があると考えている。

またこうした機能は、コンピュータを含む機械全般を対象に拡張できるため、エージェントを世界中持ち歩き、空港での座席の予約やレストランでの清算をエージェントとの日本語の対話で済ませたり、ナビゲーションシステムとエージェントを連動させて、場所に応じたタイムリーな処理や情報の提供をさせるといったことも可能になると考える。

他にも、エージェントのマルチモダリティを利用して、視覚や聴覚に障害をお持ちの方に障害のないモダリティに情報を加工して提供したり、体の不自由な方にエージェントの家電や機器の代行操作機能を活用していただくといった利用法も考えられよう。

さらには、エージェントをヒューマノイドロボットなどに乗り移らせ、物理的な作業をさせることも考え得る。エージェントを電子的に持ち歩き、必要に応じて、ロボットに乗り移らせて作業をさせる（物理的な「身体」は持ち歩かない）といった利用法は有用かつ現実的であろう。

5 まとめ

本稿では、前半で人間との共生を目指すシステムの「視覚」にはいかなる役割が求められるかについて考察し、後半で筆者らの最近の取り組みを述べた。

今後も科学技術は進展し、社会はますます高度化・複雑化すると思われる。そうした社会において、高齢者を含む誰もが安心・快適に最先端の科学技術（情報技術）の恩恵を受けられるようにすることは、世界一の長寿を誇る我が国の大きな課題であろう。

今後さらに「人」に関する総合的な研究が進み、そこで得られた知見から人と親和性の高い知的な機械情報システムが構築されて、様々な場面で私たちを豊かにサポートしてくれる日が来ることを願ってやまない。

【謝辞】本研究は、経済産業省リアルワールド・コンピューティング（RWC）プログラムの一環として進めたものである。また本稿で述べたマルチモーダル応用システムの構築にあたっては、情報処理振興事業協会「未踏ソフトウエア創造事業」から援助を受けた。関係各位に感謝する。

参考文献

[1] O.Hasegawa, K.Itou, T.Kurita, S.Hayamizu, K.Tanaka, K.Yamamoto and N.Otsu : “Active Agent Oriented Multimodal Interface System”, Proc. IJCAI-95, pp.82 - 87, 1995.

[2] 長谷川, 坂上, 速水：実世界視覚情報を対話的に学習・管理する人間型ソフトウエアロボット, 信学論 D-II, vol.J82-D-II, no.10, pp.1666-1674, Oct. 1999

[3] Bryan Adams, Cynthia Breazeal, Rodney Brooks, and Brian Scassellati. Humanoid Robots: A New Kind of Tool, IEEE Intelligent Systems, Vol. 15, No. 4, pp. 25-31, July/August, 2000.

[4] The Robot Learning Laboratory, CMU, http:// www.cs.cmu.edu/ rll/ index.html

[5] Y.Matsusaka, T.Tojo, S.Kubota, K.Furukawa, D.Tamiya, K.Hayata, Y.Nakano and T.Kobayashi, Multi-person conversation via multi-modal interface - A robot who communicate with multi-user-, Proc. Eurospeech 99, vol.4, pp. 1723-1726, Sep., 1999

[6] 石塚 満： “マルチモーダル擬人化エージェントシステム”， システム / 制御 / 情報， Vol.44, No.3, pp.128-135 (2000.3)

[7] 長谷川, 森島, 金子：「顔」の情報処理, 電子情報通信学会論文誌 (A),vol.J80-A, no.8, pp.1231-1249, Aug. 1997

[8] オリバー・サックス：火星の人類学者 - 精神科医と７人の奇妙な患者 -、ハヤカワ文庫 NF、 2001

[9] 川人光男：脳の計算理論、産業図書、 1996

[10] 川人、銅谷、春野：ヒト知性の計算神経科学（連載：第１回～第５回）、科学、岩波書店、 2000 ～ 2001

[11] 長谷川：実世界知能を創る、第４回知能情報メディアシンポジウム論文集, pp.69-76, 1998.

[12] 長谷川：マルチモーダル研究の現状と展望、電子情報通信学会 パターン認識とメディア理解研究会 技術報告、PRMU2000-106, pp.47-52, 2000

[13] 長谷川修： “生体の視覚に学ぶコンピュータビジョン”, 情報処理, J.of IPSJ, vol.39, no.2, pp.133-138, 1998

[14] J.Weng, J.McClelland, A.Pentland, O.Sporns, I.Stockman, M.Sur, and E.Thelen: Autonomous Mental Development by Robots and Animals, Science, January 26; 291: 599-600, 2001.

[15] The 2nd International Conference on Development and Learning (ICDL'02): http:// www.egr.msu.edu/ icdl02/

[16] Y.Komura, R.Tamura, T.Uwano, H.Nishijo, K.Kaga and T.Ono : Retrospective and prospective coding for predicted reward in the sensory thalamus, Nature, vol.412: pp.546-549, 2001

[17] 塚田稔：海馬神経回路の長期増強と学習則、丹治・吉沢編、脳の高次機能、 pp.187-208, 朝倉書店、 2001

[18] 森田昌彦：記憶と思考の神経回路モデル、丹治・吉沢編、脳の高次機能、 pp.211-229, 朝倉書店、 2001

[19] I. Tsuda: Towards an interpretation of dynamic neural activity in terms of chaotic dynamical systems, Behavioral and Brain Sciences 24(4) (2001).

[20] Gallese V. and Goldman A.: Mirror neurons and the simulation theory of mind-reading., Trends in Cognitive Sciences 2, 493-500, 1998

[21] 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-A, No.8, pp.1215-1230, 1997

[22] Deb Roy : “Learning from Sights and Sounds: A Computational Model.”, Ph.D. Thesis, MIT Media Laboratory. 1999.

[23] ETL 顔 CG 公開のページ: http://www.etl.go.jp/etl/gazo/CGtool/

[24] 無藤隆：赤ん坊から見た世界, 講談社現代新書, 1994

[25] 赤穂昭太郎, 速水悟, 長谷川修, 吉村隆, 麻生英樹: “EM 法を用いた複数情報源からの概念獲得”, 信学論, Vol.J80-A No.9 pp.1546-1553, 1997