ユーザ支援のための携帯型マルチモーダル対話エージェントの開発

Development of a multimodal portable agent and its future

長谷川 修
Osamu HASEGAWA

産業技術総合研究所 脳神経情報部門
（〒305-8568 つくば市梅園 1-1-1 E-mail: o.hasegawa@aist.go.jp）

Abstract　In this paper, we introduce a multimodal portable agent which supports users at any time and any place. The agent is installed on a notebook-typed computer and is displayed as a human-like CG (computer graphics) character. The agent has functions of visual recognition, speech understanding and speech synthesis. In the other part of the paper, we discuss future directions of the multimodal interaction research and its future.

1　背景

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。

これまでに試作したロボットは、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いる。

ロボットの基本性能としては、顔表情と微妙な視線、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[11]。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[12]の形成が可能であることを意味している[2]。

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[13]）させるものである[10]。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1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。

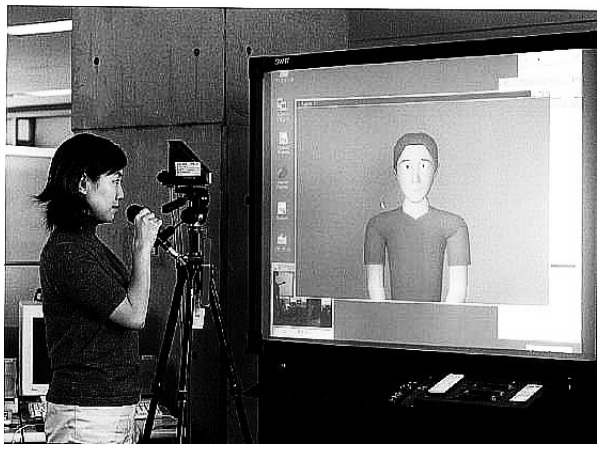

図1：人間型ソフトウエアロボットのプロトタイプと人間の対話の様子

2　目的

本研究開発の目的は、上記の人間型ソフトウエアロボットをベースに、個人が気軽に利用できる携帯型の対話エージェントを具体的に構築することにある。

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。

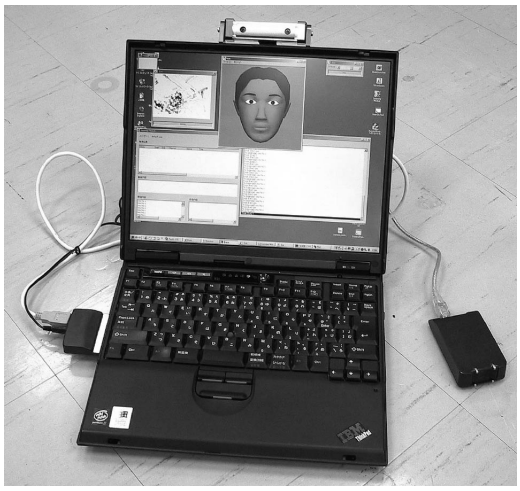

図２：マルチモーダル携帯エージェントの試作機

図３：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。）

3　開発システム

開発した携帯型エージェントシステムは通常は図１に示す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３）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４）。これにより認識率が格段に向上した。

図４：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。）

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た（図５）。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。

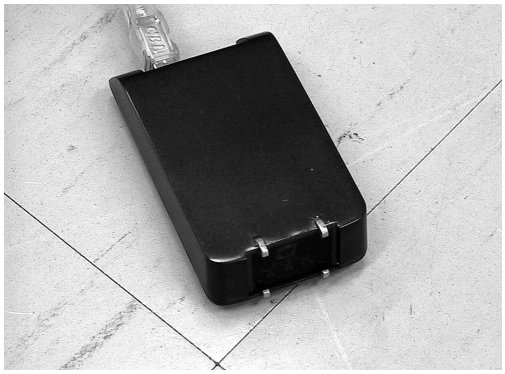

図５：通信デバイス（汎用リモコン。エージェントと家電などとの通信に利用している。制御対象の識別信号を受け、それに対応した制御信号を発信する。）

図６：マルチモーダル携帯エージェントの処理画面例

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。

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。

なお「背景」に述べた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り、本開発の重要な先行／導入研究となっている。関係各位に感謝する。

4　協力企業及び機関

- メディアドライブ株式会社
- 株式会社 ビュープラス
- 株式会社 ハル・コーポレーション

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] 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

[9] 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.

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

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

[12] 無藤隆：赤ん坊から見た世界, 講談社現代新書, 1994

[13] 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