コンピュータビジョンと
イメージメディア 100－4
（1996．5．24）

視覚情報を対話的に学習するマルチモーダル擬人化エージェント

長谷川修*，坂上勝彦，伊藤克亘，栗田多喜夫
速水悟，田中和世，大津展之

電子技術総合研究所
〒305　茨城県つくば市梅園 1-1-4

*hasegawa@etl.go.jp

本稿では、エージェント型のマルチモーダル対話のプロトタイプシステムについて述べる。本システムは、視覚、聴覚、発話機能を有し、モニタ上には擬人化 CG エージェントを表示している。ユーザはこのエージェントとの対話を通じ、視覚情報を効率的にシステムに学習させることが可能な他、豊かな表情と腕・胴を用いた様々なジェスチャの表出が可能となっている。

A Multimodal Anthropomorphic Agent which Learns Visual Information Through Interactions

O.Hasegawa, K.Sakaue, K.Itou, T.Kurita
S.Hayamizu, K.Tanaka and N.Otsu

Electrotechnical Laboratory
1-1-4, Umezono, Tsukuba, Ibaraki, 305 Japan

hasegawa@etl.go.jp

This paper presents a maltimodal interface system with functions of seeing, hearing, speaking and showing. The system synthesizes an anthropomorphic "agent" with body and arms on its display. That is, users can communicate with the system via this agent. Additionally, users can make the system learn visual information shown to the system.

1 はじめに

近年、人の有する複数のコミュニケーションモードの特性を理解し、これを活用して人とコンピュータのスムーズな対話を目指す「マルチモーダル対話(Multi-Modal Interaction)」の研究の重要性が認められ[1]、大学や各研究機関、RWCプロジェクト[2]などにおいて活発に推進されている。

マルチモーダル対話の研究では、音声／言語によるverbalな情報と視覚によるnon-verbalな情報との有機的結合による統合的認識、またその人工的な生成が主要なテーマとされており、既にその実現を目指した試作システムが各所から報告されている[5, 6, 8]。一般にこれらのシステムは、入力系として音声や画像の認識機能を備える他、出力系としてコンピュータのモニタ上にエージェントと呼ばれる仮想の「人物」を生成してこれとユーザに対話させる構成が取られることが多い。

しかしこれまでに提案されたシステムは、音声対話システムをベースに拡張されたものが多く、画像の認識や合成の機能は音声対話の補助的な役割を果たすに留まっているように見受けられる。

そこで筆者らは、マルチモーダル対話システムにおいても画像認識／合成技術を積極的に活用する試みを進めており、95年にエージェントがユーザを識別し自発的に対話を始めるシステムを提案した[8]が、本稿ではこれに改良を加えたので報告する。

改良の第一は、「対話を通じた視覚情報の学習機能」の付加である。

エージェントがユーザを識別したり、画像認識の結果を用いて対話を行なうマルチモーダル対話システムを想定すると、新たなユーザの登録や、それまでエージェントが見たことのない物体を見せてその名称を教えたりする状況が考えられる。95年のシステムでは、あらかじめユーザの顔画像を学習させて学習済のユーザの識別を行う構成としたが、本システムでは対話を通じての新しいユーザの登録や、任意の物体をシステムに見せることによる登録も可能となった。即ち、これらの拡張によって

- 新たなユーザの対話を通じた登録
- システムを頻繁に使用する人のみを登録することによる、画像や音声の認識／識別精度の向上
- ユーザの求めに応じた、物体の視覚機能による登録

が新たに可能となった。この視覚学習機能は、大津・栗田らの画像認識アルゴリズム[9, 10]の持つ特長を活用して実現したものであり、顔や物体の学習・認識を全て同一のプログラムにより行なうことができる。

改良の第二は、「エージェントからのより豊かな視覚的表現」が可能になった点である。

本システムでは、エージェントに頭（表情）の他に上半身（胴と両腕）を加え、対話の流れに応じた表情とジェスチャの表出を実現している。エージェントに腕を加えることは、顔のみを有する場合に比較して極めて大きな表現能力の拡張となっている。これは人間同士で対話をする際の状況を顧みれば明らかであろう。

さらに、エージェントに腕があることで、エージェントが仮想環境や実環境内の物体や人間を指し示すことが可能となる。これによって、そうした物体を扱ったり、複数の人間がいる状況で特定の人物を指示したりなどのように、エージェントの実行できるタスクは格段に広がる。

また、大きさや位置の情報のように、音声言語では表現の困難なものも表現できるようになるなど、対話におけるノンバーバルな表現能力が、顔のみの場合に比べて拡大している。

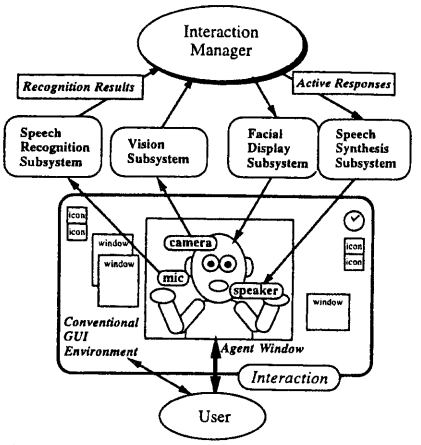

図1 本プロトタイプシステムの構成

2 システムの概要

2.1 システムの構成

本システムは、画像の認識・合成と音声の認識のためのワークステーションを数台の他、音声合成器、画像入力用CCDカメラ、マイクなどからなる構成となっている。これらのハードウエアは全て市販のものであり、特殊仕様の装置は用いていない。図1に、本システムの構成を示す。

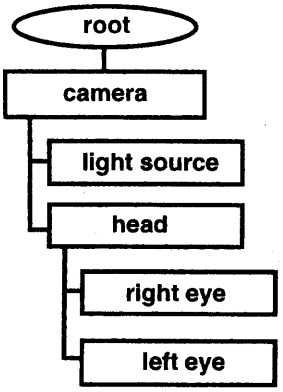

図2 頭部生成のための座標系の構成

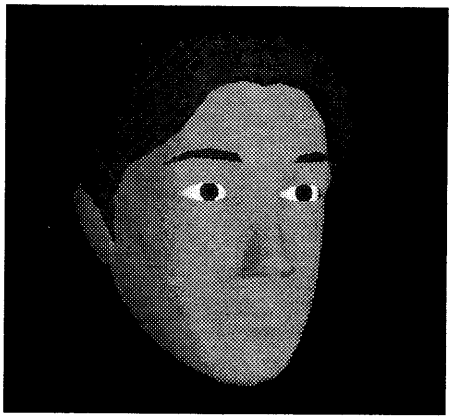
図3 本システムで用いたエージェントの頭部

2.2 エージェントの合成

2.2.1 エージェントの利用

本システムでは、若い男性の上半身を模したエージェントを利用したが、対話システムにおけるこのようなエージェントの利用は、主として

- 機械に対するユーザの「親しみやすさ」「話しかけやすさ」の向上
- システムの稼働状態の分かりやすい提示

のために有効であることを確認している [3, 4]。

また本システムのエージェントでは、必要以上に外見のリアリティを高める処理は敢えてせず、誰にでも一見して機械による合成画像であることが分かるようにした [?]。これは、エージェントとの対話は実在する人物との対話とは質的に異なることをユーザに明示する必要があるとの認識に基づいている。

2.2.2 エージェント・モデル

エージェント像は、3次元的に配置したポリゴン群をそれぞれ適切な色でレンダリングして生成しており、ポリゴン数は頭の部分で約 1700 個、両腕を含む胴体部分で約 3000 個である。このモデルは市販のモデラを用いて独自に作成したものである。

エージェントの頭部は、平常状態、喜び、驚き、悲しみ（困惑）の各表情と、うなずき、瞬きなどの挙動の表出が可能となっている。さらに、左右の眼球をそれぞれ独立に制御することによって、「輻輳」の表現も可能である。

図2に頭部生成のために設定した座標系の構成を示す。図中、上部に位置する座標系は下部の座標系を内包することを示している。これにより、「頭が回転する際には眼球も一緒に回転する」、「カメラ（視点）位置を移動してエージェントの側頭部を見る」などの処理を実現している。

表情の表出は、各表情のポリゴンのパターンを予め用意し、ある表情とこれから表出しようとする表情のパターンの差を、nステップで補間することにより行なっている。この操作は顔全体に対し一律に行なう以外にも、眉、まぶた、口に対して選択的に行なうことができ、例えば口だけを笑わせるなどの処理も可能である。

また本エージェントは、顔の向きによらず眼球を端末画面前方に向けることにより、ユーザと常に「視線の一致」を保ち続けることが可能となっている。

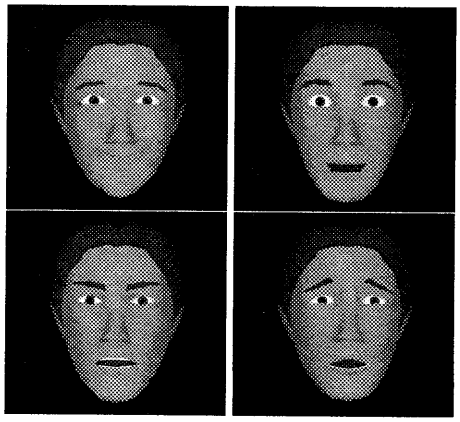
図4 エージェントの各種表情の合成例
（微笑み、驚き、怒り、悲しみ／困惑）

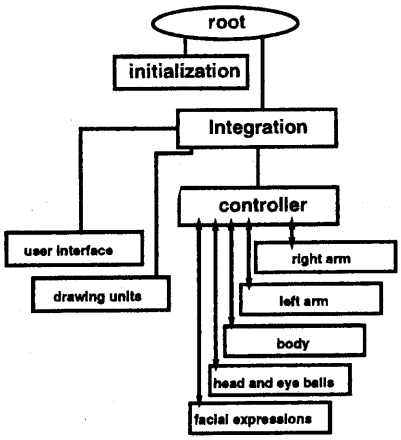

図5 全身生成のための座標群の構成

一般に視線の一致は、Face-to-Face の対話においては重要な要素であり、これは人物型エージェントとの対話においても同様である。

図3に本システムで用いたエージェントの頭部を示し、図4にエージェントが表出する表情の例を示す。図の左上部から時計回りに、微笑み、驚き、怒り、悲しみ（困惑）である。

エージェントの体の部分には、胴と両腕までを持たせている。図5に本エージェント全体の座標系の構成（概要）を示すが、上部に位置する座標系は下部の座標系を内包することを示している。

本エージェント全体で設定した座標系は合計で31、自由度は55である。

本システムにおけるエージェントの合成速度はSGI Indigo2 Maximum Impact を用いて毎秒20フレーム以上となっている。これは、動画としての知覚には充分な速度である。

2.2.3 エージェントの実装

エージェントの挙動の記述とその合成にはC++, OpenGL, GLUT を用いているが、本モジュールは主として、各ポリゴンの動きを算定する挙動演算部とポリゴンを描画するCG生成部とから構成される。

挙動演算部では、入力に対するエージェントの応答（挙動）を生成するための、頭や胴の回転角、腕の制御のための逆キネマティクスの演算を行なっている。

CG生成部では、オブジェクトとして tree 状に配置されたエージェントの各パーツ（眼球、頭、胴体、腕、指など）を、演算部の演算結果に基づいて実際に描画し、表示する。図6に、このようにして生成されたエージェントの表示例を示す。

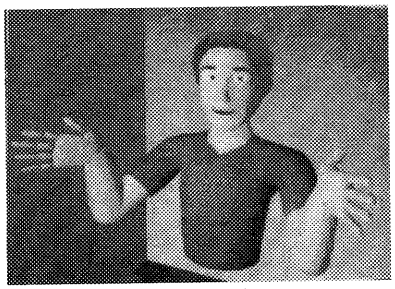

図6 エージェントの生成例

2.3 画像認識モジュール

2.3.1 概要

画像の認識は、大津・栗田らの入力濃淡画像 (90x60 pixels) の高次局所自己相関特徴と判別分析に基づく学習を組み合わせた手法により行っている [9, 10]。

本システムにおける画像の学習・認識速度は、カメラからワークステーションへの画像の取り込みと表示を含め、秒間約5フレームであった。また画像の認識能力では、最近の顔画像を用いた研究で116人分の顔が最高98%の精度で識別されている [11]。

2.3.2 本システムにおける利用

本画像認識手法は

- 認識対象を（顔などに）限定しない
- 認識対象を画像中の位置に対して不変に認識することが可能
- 学習に要する学習時間は短時間で済む

など、インタフェースシステムに適用する上で好ましい特長を多く備えている。そこで本システムでは、このアルゴリズムによって、新しいユーザの登録（顔画像の学習）と、ユーザからの要求に応じてのカメラの前の対象物の学習を同時に実現した。

これは具体的には、対象の学習時に画像から抽出される自己相関特徴を認識対象のサンプル毎に保存し（1対象につき高々数キロバイト）、識別対象が増える度に判別特徴空間を再構成することにより行なっている。

なお、栗田は別途にパターン認識のための平均マハラノビス汎距離による逐次更新アルゴリズムを提

案している [13] が、ここでは処理の簡素化を図り、上記の処理プロセスを採用した。

なお、ここで要する学習時間は、認識対象が20程度の場合、SGI の Indy R4400 200MHz を用いて1秒程である。

2.4 音声認識モジュール

音声認識部は、文献 [12] のシステムを基にしている。

本システムでは、まず入力されたアナログ音声波形をＡＤ変換する。次に振幅によって音声区間を推定してその区間だけベクトル量子化し、そのパラメータ列がどのような単語列であるかを、あらかじめ用意した音韻の標準パターンや辞書、文法を利用して推定している。

発話の理解率は、不特定話者 40 人による 183 発話で 84.2% である。なおここでいう理解率には、対話を進める上で問題にならない認識誤りは正解として含めている。

2.5 音声合成モジュール

音声合成には、市販の規則音声合成器を利用している。本システムでは、この合成器用に開発された市販のソフトウエアを用い、発話文（かな漢字混じり文）をアクセント情報などを付与した読みの情報に変換し、合成器に送っている。

ただし、語のアクセントパターンなどはほぼ正しく発声することは可能だが、声の表情の表現や文脈による強調などは行っていない。なお、今回のプロトタイプでは、定型文を十数種あらかじめ用意し、状況に応じて使い分けて利用した。

2.6 統合管理モジュール

統合管理部では、システムへの入力に応じて各種のタスクを起動する。具体的には、音声や画像の認識結果に応じて実時間でエージェントの表情と挙動を変化させたり、応答文の生成と合成などを行っている。

また、認識対象が人物か物体かの情報や、また物体の所有関係に関する情報の保持／管理もここで行なっている。

3 エージェントとの対話

3.1 準備

まずエージェントが表示されるモニタの上部にカメラを設置し、ユーザの顔画像と名前、また背景の画像をシステムに学習させる。なお、背景と照明は通常のオフィス環境である。音声まわりでは、ユーザに個別の初期設定を行う必要はない。

3.2 対話例

以下に対話例を示す。ここでＡはエージェント、●●は登録済みのユーザ、○○は未登録のユーザであり、＜＞内はエージェントの挙動である。

（システムの前に●●登場）
A　　：こんにちは、●●さん。
＜おじぎで挨拶＞

●●　：こんにちは。今日は何日ですか？
A　　：今日は７日です。

●●　：これを覚えて下さい。
（覚えさせたいものをカメラの前に差し出す）
A　　：わかりました。学習を始めます。
＜じっくり見る仕草＞
...終了しました。
その名前と持ち主を教えて下さい。
●●　：（名前☆☆、持ち主××を入力）
A　　：　××さんの☆☆を登録しました。

●●　：これは何ですか？
A　　：それは××さんの☆☆です。
●●　：ありがとう。
A　　：どういたしまして。
＜笑顔で会釈＞

●●　：新たな人を登録します。
A　　：了解しました。どうぞお座り下さい。
＜システム前の椅子を指し示す＞
（登録される人と入れ替わる）
よろしいですか？
○○　：はい。
A　　：学習を開始します。
＜じっくり見る仕草＞
終了しました。お名前を教えて下さい。
○○　：（名前を○○と入力）
A　　：○○さんを登録しました。
○○　：じゃあね。
A　　：さようなら、○○さん。
＜おじぎで挨拶＞

3.3 考察

このように、本システムでは視覚入力（画像認識）機能を活用し、対話の相手を能動的に理解しての対話と、またユーザの求めに応じた認識対象のその場での学習を実現した。一般に視覚情報は言語による記述が困難な場合が多いため、画像情報をそのまま学習させることができれば、有効に活用できる場面は多いと考えられる。

さらに、エージェントのジェスチャを活用することにより、システムの稼働状態のユーザへのフィードバックを、顔のみの場合に比較してより適切に行なうことができることを確認した。

ただし、現段階では新たな認識対象の名称をキーボード入力により行なっているため、ここを音声入力などに変更するなどの改善を検討している。

3.4 まとめ

本稿では、我々が構築を進めているエージェント型マルチモーダル対話システムの幾つかの改良点について述べた。

我々は今後も、ユーザの発話と挙動からユーザの意図を汲みとったり、ユーザとの対話を通して知識を自ら獲得・学習する対話システムの実現を目指して研究を進める予定である。

謝辞

本研究は、RWCプロジェクトの一環として行われたものである。関係各位、並びに電総研新情報計画室の皆様に感謝する。

付記

本稿で述べた CG エージェントは、通産省工業技術院研究情報基盤整備事業 (Research Information Network Grand Challenge Program = RING Program) の一環として開発したものであり、段階的に無償公開することを予定している[1]。

参考文献

[1] 黒川隆夫：ノンバーバルインタフェース, オーム社, 1994

[2] RWC 情報統合ワークショップ予稿集'96, 新情報処理開発機構, 1996

[3] K.Itou et al.:"Collecting and Analyzing Nonverbal Elements for Maintenance of Dialog Using a Wizard of Oz Simulation", Proc. Int'l Conf. on Spoken Language Processing, pp.907-910, 1996

[4] 長谷川修, 伊藤克亘, 秋葉友良, 速水悟, 田中和世, 山本和彦. :" 音声対話システム利用時のユーザの振舞いの解析"·第 10 回ヒューマン・インタフェース・シンポジウム論文集, pp.631-636, 1994.

[5] Riecken D. ed.:"Intelligent Agents", COMMUNICATION OF THE ACM, Vol.37, No.7, 1994

[6] 土肥浩, 石塚満："WWW/Mosaic と結合した自然感の高い擬人化エージェントインタフェース", 信学論 D-II, Vol.J79-D-II, No.4, pp.585-591, 1996

[7] 平川正人、安村通晃編：" ビジュアルインタフェース", 共立出版, bit 別冊, 1996

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

[9] N. Otsu and T. Kurita, "A new scheme for practical flexible and intelligent vision systems," Proceedings of IAPR Workshop on Computer Vision –Special Hardware and Industrical Applications–, Tokyo, October 1988, pp. 431-435, 1988.

[10] Kurita T., Otsu N. and Sato T. :"A Face Recognition Method Using Higher Order Local Autocorrelation And Multivariate Analysis", Proc. of 11th ICPR, The Hague. Vol . II, pp.213-216, 1992

[11] F.Goudail et al:"Fast Face Recognition Method Using High Order Auto- correlations", Proc.of IJCNN, pp.1297-1300, 1993

[12] Itou K. et al.:'System design, data collection and evaluation of a speech dialogue system", IEICE Trans, INF.&SYST., Vol. E76-D. No.1, pp.121-127, 1993

[13] 栗田多喜夫：" 平均マハラノビス汎距離によるパターン認識のための逐次更新アルゴリズム" , 信学会秋期全国大会予稿集, D-322, 1992

[1] http://www.etl.go.jp/aist/www_j/aistring.html