実空間中の人物や物体を認識して対話するマルチモーダル擬人化エージェント* 3B−8

長谷川修, 速水悟, 坂上勝彦
電子技術総合研究所

1 はじめに

インタフェース・エージェントとしての擬人化エージェントとユーザ（実世界）のスムーズなインタラクションのために、視覚情報は極めて有用と考えられる。そこで筆者らは、先にユーザや物体の画像情報を対話を通じてオンラインで学習し、学習後は実時間でそれらを認識するシステムを構築した [1, 2] が、今回これに改良を加えたので報告する。

具体的には、(1) 計算機から制御可能な雲台付きカメラを導入し、エージェントの“視野”を拡げた。(2) 複数の対象物を並行的に認識可能とした。(3) 認識結果は“何が、どこに、何時から、何時まで、何回”現れたかという履歴として統合管理部（後述）で管理し、対話時に利用可能とした。(4) 予め顔画像と性別を学習させてある人物との対話時には、画像認識結果に基づき、その人の性別にあわせた音声認識用音韻モデルに自動的に切替えるようにした。(5) 認識した対象の実空間中の位置に関する問い合わせに対しては、エージェントの発話・視線・指さしジェスチャを併用し、ユーザが直観的に把握しやすい形式で応答するようにした（図1）。

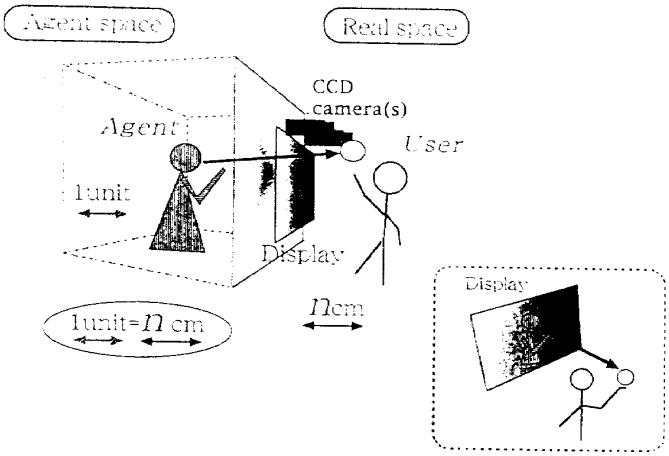

図1：エージェントは実空間（実世界）中の人物や物体を認識し、それらの位置を「〜はここ／そこ／あちらです」という発話、視線の向き、指さし動作を併用してユーザに示すことができる。

2 システムの概要

2.1 ハードウエア構成

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3台の他、画像入力用雲台付きカメラ、音声入力用マイク、音声合成器などからなる構成となっている。これら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。

2.2 エージェント像

エージェント像は人の上半身の姿を有し、ポリゴン数は頭の部分で約 1700、両腕を含む胴体部分で約 3000 である。本エージェントは、喜び、驚き、悲しみ（困惑）などの表情に加え、微妙な輻輳（視線）の表現や、指さしなどの十数種のジェスチャの表出が可能である（図2）[1]。

図2：指先の注視（本システムではこうしたエージェントの視線や指さしのジェスチャも活用する。）

2.3 音声の認識と合成

ユーザの発話を認識する音声の認識部は、文献 [4] のシステムを基にしている。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。またエージェントの発話の合成には市販の音声合成器を利用しており、約 20 種類の定型文を状況に応じて使い分ける構成としている。

2.4 画像の認識

2.4.1 背景と認識対象の学習と認識

本システムにおける画像の学習・認識手法の概要を以下に示す。本手法は画像からの高次局所自己相関特徴の抽出とその判別分析による学習という枠組に基づく [5]。

[1] 背景クラスの構成

1. システムからユーザの側に向けて設置したカメラを 0, ± a1,..., ± an(rad) 回転させ、背景となるシステム周辺の画像を入力する。
2. 入力した背景画像を合成し、背景のパノラマ画像を構成する。
3. パノラマ画像を α 個の小領域 R に矩形分割し、各 R とその近傍から q 枚の画像を得る。これを各々 q 個のサンプルからなる α 個の背景クラスとする。
4. 上記の $\alpha \times q$ 枚の画像から各々高次局所自己相関特徴を抽出し、各クラスの基本特徴データとする。

[2] 認識対象クラスの構成

5. 上記カメラのオンライン入力画像に上記 R と同サイズの矩形を表示し、ユーザに白枠で提示する。
6. ユーザはその枠内に認識させたい対象（自らの顔や物体）を提示し、システムはそれらの様々な“見え”の画像を q 枚入力する。
7. 以上の過程を認識対象の数（β 個とする）と同数回繰り返し、各々 q 個のサンプルからなる β 個の認識対象クラスとする。

---

*“ A multi-modal interactive agent which recognizes visual objects and users in real-space”, Osamu HASEGAWA, Satoru HAYAMIZU and Katsuhiko SAKAUE, Electrotechnical Laboratory

[1] 本エージェントの生成プログラムは無償公開している [3]。

8. 上記の $\beta \times q$ 枚の画像から高次局所自己相関特徴を抽出し、各クラスの基本特徴データとする。

[3] 判別空間の構成

9. 以上で抽出した $(\alpha+\beta)$ クラス分の特徴データを判別分析する。これにより認識対象を“抽出”するための判別空間を構成する。
10. 認識対象クラスの基本特徴データをのみ判別分析する。これにより認識対象を“識別”するための判別空間を構成する。

以上で「学習」の過程が完了する。

[4] 認識対象の探索と認識

11. システムが起動されると、システムは入力画像をRと同サイズのウインドウでサーチし、高次局所自己相関特徴を抽出する。
12. まず抽出データを“抽出用”判別空間に射影し、認識対象であるかどうかを判別する。ここで認識対象と判別された場合に限り、サーチウインドウの画像上での位置を矩形で表示して13に進む。
13. 抽出データを“識別用”判別空間に射影し、いずれの対象であるかを判別し、結果を統合管理部に送って11に戻る。

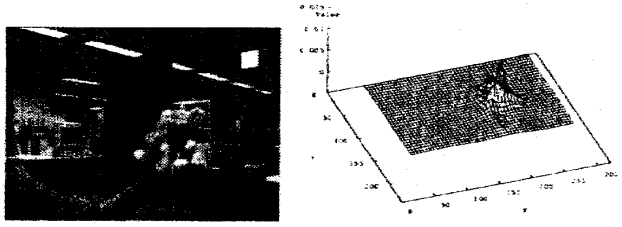

図3：対象物の探索・認識結果1（左：入力画像、右：認識対象クラスとの類似度マップ）

図4：対象物の探索・認識結果2
（4つの対象物の並行認識例）

2.4.2 認識対象の探索・認識実験と結果

本手法は対象に依存しない特徴を用いるため、人の顔や物体を統一的に扱うことができる他、様々な特性を有している[5, 6, 7, 8, 9]。また、40クラス、1200サンプル程度の学習ならば数分で完了するが、これは例えばニューラルネットなどによる学習に比較して格段に短い学習時間といえよう。

本システムにおける探索・認識実験の結果例を図3、4に示す。図3右図のメッシュの交点はウインドウRがサーチした入力画像上の位置（正確にはRの右上の角の位置）を示し、縦軸はその位置の画像情報の判別結果を判別距離の逆数で示したものである。この例では、左図に示す「きつねの人形」（認識対象の一つ）が正しく認識（判別）されるとともに、右図のピーク位置によってその画像上の位置も正しく検出されている。

図4は画像中に含まれる4つの認識対象を同時並行的に認識した結果例である。

実験の結果、図4に示すようなオフィス環境でカメラと認識対象との距離を約0.5～2mとした場合、10種類（クラス）の人の顔や物体の判別・認識率は93.3%であった。

2.5 統合管理部とエージェントの挙動

統合管理部はエージェントの“脳”に相当するシステムの中枢であり、上記の画像の認識結果もここに逐次送られて来るが、本システムではこのデータを“何が、どこに、何時から、何時まで、何回”現れたかの履歴として記録・活用することとした。この結果、システムはユーザの「～さんはここに来ましたか。いつ来ましたか。」といった問いや「～はどこにありますか。」といった問いにも応答できるようになった[2]。さらに、後者の問いに対し、エージェントに「～はここ／そこ／あちらです。」という発話と視線／指さし動作を併用して応答させたところ、10人の被験者による実験の結果、直観的な理解を得やすく、こうした被験者の空間的な注意／注目位置の誘導に有効との結論を得た。

図5：ユーザとエージェントの対話中の様子

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。関係各位に感謝する。

参考文献

[1] O.Hasegawa, et al.: “Active Agent Oriented Multimodal Interface System”, Proc.IJCAI-95, pp.82-87, 1995.
[2] 長谷川他：“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”, 情処 CVIM100-4, pp.33-38, 1996
[3] ETL CG Tool のページ：http://www.etl.go.jp/etl/gazo/CGtool/
[4] 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
[5] 大津：パターン認識における特徴抽出に関する数理的研究, 電総研報告, 第818号, 1981
[6] 長谷川, 栗田, 坂上：“学習によるシーン理解の提案とその基礎実験”, 第3回画像センシングシンポ, C-18, pp.129-132, 1997
[7] 曽根, 長谷川, 坂上：“部分画像からの物体の認識と切り出し手法の提案”, 情処56回全大, 掲載予定, 1998-3
[8] Yoda I., Sakaue K.：“Utilization of Stereo Disparity and Optical Flow Information for Human Interaction”, Proc. 6th ICCV, pp.1109-1114, 1998
[9] Raytchev B., Hasegawa O., Otsu N.：“Gesture Recognition By Geometrical Statistical Feature Extraction And Discriminant Analysis”, 情処56回全大, 掲載予定, 1998-3

[2] 実験結果に示すように現状では認識対象までの距離に制限がある。カメラのズーム機能の併用などにより解決を試みている。