様式 C-19

科学研究費補助金研究成果報告書

平成２１年５月１９日現在

研究種目：若手研究（B)
研究期間：2007～2008
課題番号：19700179
研究課題名（和文）　パターン認識器統合フレームワークとデータ収集に関する研究

研究課題名（英文）　Study on Pattern Recognition Integration Framework and Data Collection

研究代表者
藤江　真也（FUJIE, Shinya）
早稲田大学・高等研究所・助教
研究者番号：00367062

研究成果の概要：
音声認識や画像認識などのパターン認識器をシステムに組み込む為のフレームワークの構築，及び統合に適したパターン認識器の構築を行った．特に，音声対話システムのようなリアルタイム処理が必要となるシステムにおけるパターン認識器の構築について検討した．提案フレームワークを用いた音声対話システムを構築し，会話ロボット上に実装した．

交付額

（金額単位：円）

	直接経費	間接経費	合　計
2007 年度	**1,300,000**	**0**	**1,300,000**
2008 年度	**1,100,000**	**330,000**	**1,430,000**
年度			
年度			
年度			
総　計	**2,400,000**	**330,000**	**2,730,000**

研究分野：総合領域
科研費の分科・細目：情報学・知覚情報処理・知能ロボティクス
キーワード：パターン認識，音声認識，画像認識

１．研究開始当初の背景

音声認識や画像認識など，様々なパターン認識の研究は計算機の発達とともに盛んになり，実用化され始めていたが，各認識器が個別に実装されていたため，大規模なシステムに組み込む際の統合が難しかった．特に申請者の研究対象であるマルチモーダル音声対話システムにおいては，音声認識や顔画像処理などをリアルタイムで実行・統合し，総合的な判断を行うことで初めて人との対話が実現される．従来の認識器・および統合フレームワークは時間情報を積極的に利用しなかったため，音声対話システムのようなリアルタイムで動作するシステムには不向きであった．

２．研究の目的

音声認識や画像認識等のパターン認識器をシステムに組み込む際のフレームワーク，およびそれを用いて統合するのに適したパターン認識器を実現する．

対象とするシステムは，音声認識，画像認識を用いることが必須である音声対話システム，特に会話ロボット上の音声対話システムとする．

３．研究の方法

はじめに認識器統合のフレームワークの開発を行う．提案フレームワークは，各認識器の結果を簡易にやり取りできるようにし，各認識器が強い結びつき（例えば同一プロセスで存在するなど）を持たなくても連携が可能なようにする．

次に，画像処理システム，音声認識システムの開発を行う．これらは後に会話ロボット上の音声対話システムに，実現したフレームワークを用いて統合されることを目的として開発する．本研究では特に，対話システムに必要となる視線認識，対話相手を見つけるための上半身追跡システム，また会話システムに適した音声認識器の開発を行う．

最後に，開発した個別の認識器を提案フレームワークを用いて統合し，音声対話システムの入力として用いることでその有効性を示す．これらは音声対話ロボット上に実現される．

４．研究成果

(1) 認識フレームワークの開発

申請者が従来研究として構築したメッセージ指向型ネットワークロボットアーキテクチャ（MONEA; Message Oriented NEtwork robot Architecture）を元に，メッセージのやり取りのみによる軽量，簡易なインターフェースを備えた通信方式を採用した．

各認識器は，興味のある情報（音声認識であれば音声，画像認識であれば画像，対話システムであればそれらの結果）に容易にアクセスすることが可能になり，それぞれを有機的に統合することが可能となった．

(2) 画像認識システムの開発

①視線認識システム

会話において視線が果たす役割は様々なものがあるが，特に発話権の移動に関して重要な働きを持つ．例えば発話終了時の視線方向について考えてみる．発話終了時に相手を見ていれば，それは相手に発話権を渡すことを意図していることがくみ取れるし，別の場所を見ていれば，次の内容を考え中で，しばらく自分が喋ることをアピールしていると考えられる．このように，音声対話システムにおいては対話相手である人が，システムを見ているか見ていないかを認識することが重要となる．

従来はこの二者を静的な画像処理によって認識していたが，本研究では HMM(隠れマルコフモデル)を用いた動的情報の統合により，高精度に認識できる枠組みを提案し，実現した．また，リアルタイム性を損なわないため，HMM で認識した精度の高い認識結果を用いて高速・低信頼な静的パターン認識を用いた認識器の適応を行うことで，高速・高信頼な認識システムを構築した．

②上半身追跡システム

音声対話システムが対話を開始するために，対話相手となりうる人を見つけだす，あるいはその人がどのような姿勢でいるかを認識することが必要である．また，身体を用いたジェスチャ等を認識する際に，カメラの動きをキャンセルするためにも，人（胴体の輪郭）の姿勢を推定することが重要となってくる．本研究ではこれを，胴体の動きをモデル化した形状モデルを入力画像にフィッティングさせるという方法で高精度に行う手法を提案する．

胴体の形状モデルは，顔のフィッティングによく用いられる AAM（Active Appearance Models）を応用し，胴体の輪郭上に特徴点を配置し，それらの座標情報に対して主成分分析を行うことで構築した．

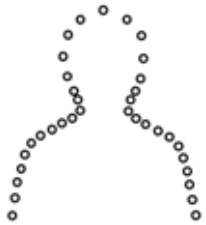

図 1 形状モデルに用いる胴体形状

フィッティング手法は AAM でも用いられている Inverse Compositional Image Alignment を用いた．この手法は入力画像をモデルとそのパラメタを用いて変換した画像と，モデルが元々持っている平均的な画像との間の誤差を最小化するパラメタを計算することで，フィッティングを行う手法である．顔をモデル化する際は平均顔を用いることができるが，本研究で対象とする上半身は服装の違いによって外観が大きくことなることから，直接用いることが出来ない．従って図 2 に示すように上半身のエッジをぼかした画像を用いることとした．ぼかす必要があるのはこの手法が画像の勾配に基づく手法だからである．

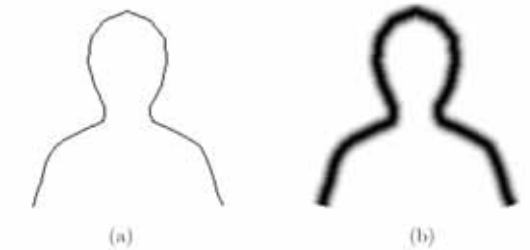

図 2．輪郭情報とした用いたエッジ画像（a）と，外観画像として用いたそれをぼかした画像(b)

構築したシステムによるフィッティング例を図 3 に示す．この図では，上半身の輪郭情報のみを用いた結果（上段）と顔モデルとの統合を行った結果（下段）を示している．上半身の輪郭モデルのみの場合はエッジ抽出の結果を信頼してフィッティングを行うため，背景のエッジに影響を受けやすい（上段右の画像でホワイトボードのエッジにフィッティング結果が引っ張られていることがわかる）．これに対し，顔モデルの統合を行うことで背景の雑音に頑健なフィッティングを行うことが可能となった．

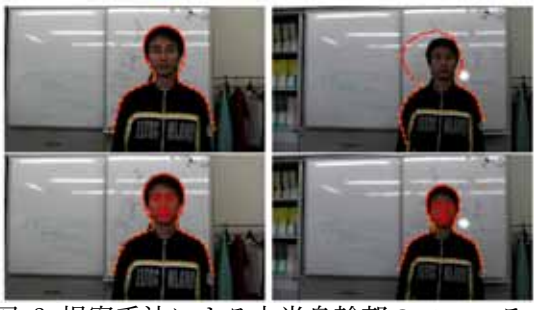

図 3 提案手法による上半身輪郭のフィッティング（下段）

提案手法の評価を行った．4 名の被験者の様々な姿勢データ（各 100 枚）を収録し，1 名分を評価データ，残りの 3 名分を学習データとして 4 分割の交差検定を行った．顔検出を行った結果から正面を向いた形状を初期形状としたフィッティングにより，収束結果がどれくらいの誤差を持つかを評価した．結果を図 4 に示す．PAACMs(Parameterized Active Appearance and Counter Models)が顔モデルと統合を行ったもの ,PACMs(Parameterized Active Counter Models)が輪郭モデルのみを用いたものである．横軸が収束結果と正解データ間の誤差で，縦軸がその誤差に収まっているデータの割合である．この結果から統合を行った場合が輪郭モデルのみを用いた場合よりもよくフィッティングしていることがわかる．

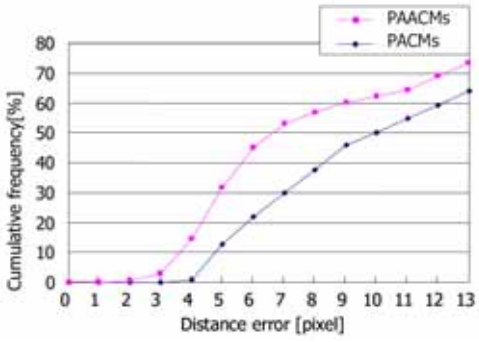

図 4 上半身輪郭モデルのフィッティング実験結果

(3) 音声認識システムの開発

音声認識システムは，入力された音声を文字列に変換するものである．音声対話システムには欠かせないシステムであるが，従来の音声認識システムが音声対話システムに適しているとは言えない．音声認識システムは，対象とする表現をどのように絞り込むかでその性能が決まる．対象を過度に広げると，曖昧性が増大し，認識誤りが多くなる．逆に対象を絞り込み過ぎると，所望の表現を認識

することが不可能になる上，会話に関係のない発話が関係ある内容と誤って認識される可能性があり，より深刻な認識誤りを起こす危険性がある．

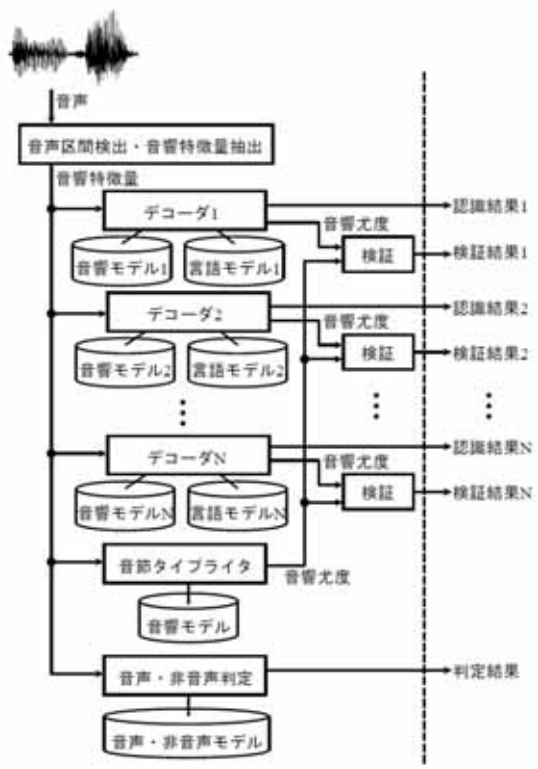

図 5　構築した音声認識器の概観

これらの問題に対し，本研究では複数の音声認識器を並列で動作させることを提案した（図 5）．会話は進行することで時々刻々と場面が変わり，それによって認識対象とすべき表現も時々刻々と変化する．さまざまな場面に応じた認識器（デコーダ）を準備し，それらを場面に応じて切り替える方法も考えられるが，場面の推定自体が誤っている可能性や，後から認識結果を振り返る際に違う場面を想定した認識結果を参照する可能性があることなどを考慮すると，一つ一つ認識を行うことよりも，並列に全ての場面を想定して認識を行った方がメリットが大きいと考えた．認識処理で最も時間のかかる音響尤度の計算部を共有することで，全体のパフォーマンスの低下を防いだ．また，認識対象としない表現を認識対象の表現と誤認識をすることを防ぐために，音響尤度ベースの認識結果棄却機構を設けた．

(4) 会話ロボットの開発

構築したフレームワークを用いて，これまでに述べたパターン認識器を統合することで，音声対話システムのフロントエンドを構築した．

構築した音声対話システムは，クイズ形式のゲームに，回答者の一人として参加する会話ロボット上に構築されたもので，司会者の発話を元にゲームの進行状況を理解しながら，他の回答者の発話を促したり，自ら回答を行うことで，ゲームを盛り上げる役割を果たす．実験の様子を図 6 に示す．

図 6 会話ロボットの実験の様子

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 3 件）

① K. Hoshiai, S. Fujie, T. Kobayashi, “Upper-body contour extraction using face and body shape variance information,” The 3rd Pacific-Rim Symposium on Image and Video Technology (PSIVT2009), Lecture Notes on Computer Science, vol.5414, pp.862-873, Jan. 2009. (査読有)

② S. Fujie, D. Watanabe, Y. Ichikawa, H. Taniyama, K. Hosoya, Y. Matsuyama, and T. Kobayashi, “Multi-modal Integration for Personalized Conversation: Towards a Humanoid in Daily Life,” Proc. Humanoids2008, vol.1, pp.617-622, Dec. 2008. (査読有)

③ T. Nakano, S. Fujie, and T. Kobayashi, “Extensible speech recognition system using Proxy-Agent,” Proc. IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU2007, vol.1, pp.601-606, Dec. 2007. (査読有)

〔学会発表〕（計 5 件）

① 藤江 真也, 渡邊 大地, 谷口 徹, 小林 哲則, “音声対話システム用音声認識器の実現と音声対話ロボットへの応用, ”人

工知能学会 言語・音声理解と対話処理研究会, SIG-SLUD-A803, pp.31-36,2009 年 3 月 13 日. （早稲田大学）
② M. Wimmer, S. Fujie, F. Stulp, T. Kobayashi, B. Radig, “An ASM fitting method based on machine learning that provides a robust parameter initialization for AAM fitting,” Proc. Int. Conf. Automatic Face and Gesture Recognition, FG2008, 2008 年 9 月 18 日. Amsterdam, Netherland.
③ 星合 和樹, 藤江 真也, 小林 哲則, “形状変化傾向を考慮した動的輪郭モデルによる人の上体輪郭へのフィッティング,” 第 11 回 画像の認識・理解シンポジウム, MIRU2008, IS-5-28, 2008 年 7 月 31 日.（軽井沢プリンスホテル）
④ 松山 洋一, 谷山 輝, 藤江 真也, 小林 哲則,“人-人コミュニケーションの活性化支援ロボットの開発,”人工知能学会 言語・音声理解と対話処理研究会, SIG-SLUD-A801, pp.15-22, 2008 年 7 月 19 日.（はこだて未来大学）
⑤ 山畠 利彦, 藤江 真也, 小林 哲則, “視線運動の離散性を用いた視線認識,”情報処理学会研究報告, 2007-CVIM-160, pp.77-82, 2007 年 9 月 3 日.（名古屋大学）

６．研究組織

(1)研究代表者

藤江　真也（FUJIE SHINYA）
早稲田大学・高等研究所・助教
研究者番号：00367062