ヒューマンインタフェース 76－7
情 報 メ デ ィ ア 33－7
（1998. 1. 29）

音声対話による画像認識支援

高橋拓弥　中西知　久野義徳　白井良明
大阪大学工学部電子制御機械工学科
565(-0871) 大阪府吹田市山田丘 2-1
Tel: 06-879-7333　Fax: 06-879-7247
{takahasi,s-nakani,kuno,shirai}@cv.mech.eng.osaka-u.ac.jp

概要

本論文では音声とジェスチャによるマルチモーダルインタフェースにより，ロボット内に存在する曖昧性を効率よく解消する手法を提案する．この曖昧性はロボットの視覚認識の失敗によって発生することが多いが，コンピュータビジョンの技術を改良するだけではこの問題は簡単には解決することはできない．我々が開発したロボットはこの視覚部分の問題点を解決するのに最も適したアドバイスを自然に人から得られるような質問を人に対して行なう．そして，人に代わってものを取ってきてくれるロボットを例にとり，人との対話の有効性を示す．

Helping Computer Vision by Spoken Dialogue

Takuya Takahashi　Satoru Nakanishi　Yoshinori Kuno　Yoshiaki Shirai
Dept. of Computer-Controlled Mechanical Systems, Osaka University
2-1, Yamadaoka, Suita-City, Osaka 565-0871, Japan
Tel: +81-6-879-7333　Fax: +81-6-879-7247
{takahasi,s-nakani,kuno,shirai}@cv.mech.eng.osaka-u.ac.jp

Abstract

This paper proposes a method of removing ambiguities in robot tasks by a multimodal interface consisting of speech and gesture. Such ambiguities often arise from failures of the robot vision system. However, it is not easy to solve this problem only by improving computer vision techniques. Thus, our robot asks a human such a question that a natural reply to it will contain helpful information to adapt the vision system for the current situation. We present a robot system that can bring the object ordered by a human by verbal and nonverbal behaviors.

1 はじめに

近年コンピュータによって制御された様々なシステムが開発されており，そのシステムの入力を司るマン・マシンインタフェイスの重要性は，ますます深まっている．人からの情報を伝達する手段として，音声やポインティングなど様々なものが使われ，またこれらを組み合わせたマルチモーダルインタフェースを使ったシステムがいくつか開発され，その有効性が確かめられている [1][2][3].

しかし，システムが複雑になればなるほど，そして人から伝えられる情報が詳細なものになればなるほど，ユーザー側の負担が増加し，システムの利便性が失われてしまう．しかも，マルチモーダルインタフェースとして，音声認識や画像認識など情報の信頼性が心配されるインタフェースを用いる場合は，システムが間違った行動をとってしまう恐れがある．特に画像認識ではシーンが複雑になればなるほど，人には簡単に認識できるものが，システムには認識できなくなるということがしばしば起こってしまう．

人工知能の分野の学習機能を用いて，これらの問題を解決することも考えられるが，最も信頼のおける解決法は人が助言を与えることであると思われる．ただし，この人からの助言は人が負担にならない範囲内に収めるようにしなければ，本来人を助けるはずのシステムが逆に人に対して重荷になってしまう．この問題は，システムが何かの障害で詰まっている時，今の状態から抜け出すには人から最低限どのような助言を得られれば良いか考え，その助言が自然な返答の形で得られるような質問を人に対してすれば解決できると考えられる．

本論文では，人に代わって物を取りに行くロボットシステムを例にとり，人とシステムとの対話の効果を示す．人は音声と身振りのマルチモーダル入力によって，対象物体の名前と位置をシステムに伝える．システムは人から入力された情報と，あらかじめ与えられている物体の情報を用いて実空間より対象物体を探索し，その物体の位置まで移動する．物体が見つからない場合は，その検出に有効な情報が人から自然に得られるような質問事項を考え，人に音声でたずねる．そして人はその質問に答え，システムはこの問いに応じて適切な画像認識処理を行なう．この人との対話を繰り返し探索物体の情報を確定する．

2 システム概要

図 1に本論文で使用するシステムの概略を示す．人は音声とジェスチャによってロボットに対して指示を送る．ロボットは人の音声とジェスチャを認識し，指示された作業を行なうのに適したプロセスを起動する．このプロセスが終る度に処理が成功したかどうか確かめる．もし処理が失敗してしまった場合は人から適切な助言が得られるような質問文を作成し，それを音声合成して人に伝える．またこの時，ロボットの動作もまた非言語メッセージとして，人に伝えられる．例えば，物体探索プロセスにおいて目標物体を見つけることができなかった場合，ロボットはカメラを左右に振る．そしてその動作は，ロボットが目標物体を見つけられず，困っているというメッセージを人に連想させる．

3 人とロボットとの対話生成

人の指示というものはたいてい，”どこで”，”何を”，”どうする”の 3 種類の要素で構成されている．人と人との対話時にはこれらのキーワードは，しばしば曖昧に表現され，又は省略される場合もある．しかし人はこの曖昧な部分や省略されている部分を推測し，指示されたことを実行することができる．また，推測できない場合には指示を出した人に，今の分かっていない状況を示し，適切な指示又は問題解決のための情報を得る．人はこのような対話を繰り返し，指示の曖昧性を取り除く操作を知らず知らずのうちに行なっていると思われる．

この考えを，人とロボットとの間の会話にも応用することを考えた．人はロボットに対して曖昧な又は省略された指示を送り，ロボットはこの指示の意味を読みとって，指示の曖昧な部分，不足部分を人に聞き返す．

ただし，ロボットの視覚認識能力が低いため，人間同志の場合よりずっと低いレベルでの曖昧性の解消を考える必要がある．例えば人が「リンゴをとって」と言った場合を考える．人間同志なら”リンゴ”についての共通な概念を持っているし，通常の場合，視野内にリンゴがあれば見つけられないこ

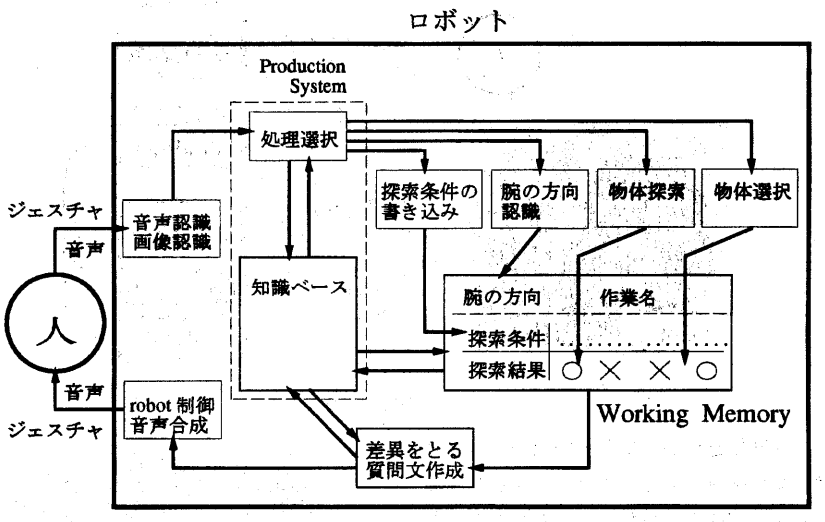

図 1: システムの概要

とはない．ところがロボットにとっては”リンゴ”の知識は例えば”赤くて丸いもの”程度にしか組み込まれておらず，しかも一般シーンの中では，そのようなものを見つけることに失敗してしまうことも多い．そこで，このようなことから起こる失敗を，人との自然なコミュニケーションにより情報を獲得し，回避していく方法を検討する．

人のコミュニケーションは言語によるものと，ジェスチャ等の非言語的なものがある．ここでは，音声対話とジェスチャやロボットの行動をコミュニケーション手段として考える．

自然なコミュニケーションを行ないながら曖昧性を解消していくためには2つの機能が必要である．一つは人の音声及びジェスチャによる指示を理解して，ロボットの行なうべき処理に翻訳する機能である．もう一つは，画像認識の結果，何が分かって何が分からないかを調べ，人に何を教えてもらえばよいかを考え，それを人から自然に引き出すことのできる質問を生成する機能である．

以上の機能を実現したシステムの構成を図1に示す．上記の2つの機能を実現するために，今何が分かっていて，何が分かっていないかを把握するための内部状態の記述スペース (Working Memory) を持っている．

例えば，「あのリンゴをとって」と人が指で位置を示しながら言ったとする．先に述べた一つ目の機能により，”リンゴ”という言葉からリンゴに関する知識を知識ベースより取りだし，WM 中の探索条件の色が赤に，形が球形にセットされる．

また，”あの”という単語が入力されたら腕の方向を認識するプロセスを起動させるというルールを知識ベース中より見つけだし，実際に認識を行なう．その結果から位置の部分にも概略の値がセットされる．個数は特に指定されなければ，デフォルトで1個としておく．このように，探索物体についての情報が得られたら，それを検出する画像処理ルーチンを起動し，その結果を探索結果に書き込む．

そして探索条件と探索結果とを比較して，差異が生じた部分があればそこを疑問点とする．この差異の取り方及び質問文の作成方法は5節で示す．

4 実験システム

実際に実験に用いたロボットの環境を図2に示す．本実験では小型移動ロボット (Real World Interface,Inc. Pioneer 1) を使用した．そして，この上には カメラ (Canon VC-C1)，ビデオ送信機と無線モデム (RS-232C 用) を搭載した．ただし，

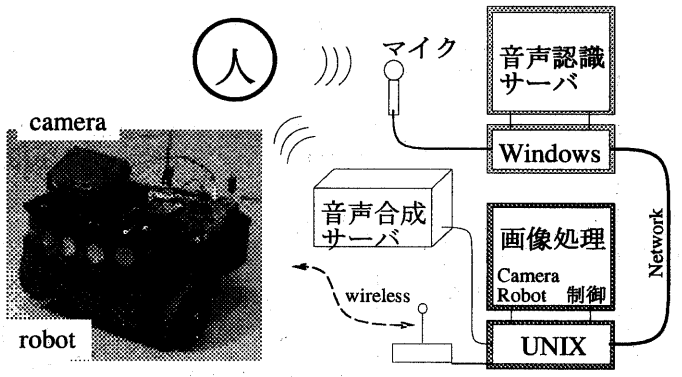

図 2: 実験システム

Pioneer 1 には物体をつかむ腕がないため，人がロボットに対して物を取ってくることを指示をした場合は，その物体の位置まで移動することができたかどうかで，作業の成功を確認することにした．音声認識処理は (株) 東芝で開発されたソフトウェア [4] を，また音声合成は NTT インテリジェントテクノロジ (株) の製品 (しゃべりん坊) を使用した．

4.1 人の身振り解析

人の身振りは，”物を指し示す” 動作のみについて考えている．そしてここでは，物を指し示している腕の方向を求める．ただし一枚の画像だけでは完全な 3 次元空間上での腕の方向を求めることはできない．しかし，人が他の人の腕の方向を見る場合も，腕の方向は完全には認識しておらず，ただ何となくそちらの方向を向くという動作をしているだけと思われる．これを考慮し，本手法でも正確な 3 次元の方向を求めるということは行なわない．連続するフレーム間の差分処理から得られた運動物体領域及び肌色領域の両者を満たす部分を腕の領域として抽出し，この領域の慣性主軸の方向を人の腕の方向としている．図 3に処理結果の例を示す．

4.2 物体探索

この処理はあらかじめ蓄えられている物体に関する知識と作業中に人から与えられた知識を用いて，対象となっている物体を画像中から探すことを行なう．用いる物体の知識として考えているものは，物体の色に関する情報と幾何学情報である．色情報を用いて画像から同一色の領域を切り出す．そ

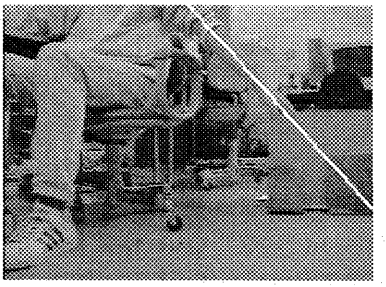

図 3: 腕の方向検出

して幾何学情報よりその切り出された領域の輪郭に直線又は楕円当てはめを行ない，その物体の形と画像上での位置を確定する．また，実際の物体の大きさも知識として与えられているため，一枚の画像からでも物体の 3 次元位置を推定することができる．図 4は画像中より赤色の領域を 1 つ切り出し，その領域の輪郭に対して楕円の当てはめを行なった結果である．画像中より 3 個のリンゴを見つけることができたことを示している．

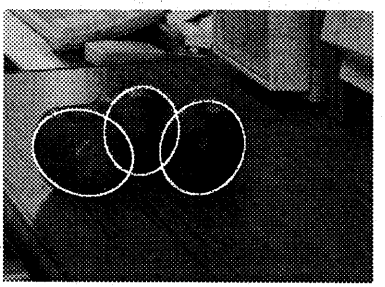

図 4: リンゴ (楕円) 検出

5 差異の取り方と対話文作成

ここでは，物体の探索条件とその探索結果の間で生じる差異とそこから対話文を作成する手順を示す．ここで言う対話文には，今ロボットが見つけた物体の特徴全てを人に対して示して，ロボットの今の状態を人に理解してもらう為の状況説明文と，その状態からどのように抜け出せば良いか人に対して聞く質問文の，二つから成り立っている．探索条件と結果の間で生じる差異として考えているものは，色，形，個数の3種類である．

これらの差異をとり，状況説明文を作るアルゴリズムを以下に示す．ここで，nは見つけた物体の個数，C_i, S_iは見つけた物体のi番目の物の色と形の名前，そして添字が0のものは探索条件を表す．

for $i = 1$ **to** n **do**
  if $(C_0 = \mathrm{NULL})$
    $(\mathrm{Color\_part})_i =$“　”
  else if $(C_0 = C_i)$
    $(\mathrm{Color\_part})_i =$“$C_0$ 色の”
  else
    $(\mathrm{Color\_part})_i =$“$C_0$ 色でない”
  if $(S_0 = \mathrm{NULL})$
    $(\mathrm{Shape\_part})_i =$“　”
  else if $(S_0 = S_i)$
    $(\mathrm{Shape\_part})_i =$“$S_0$ 形の”
  else
    $(\mathrm{Shape\_part})_i =$“$S_0$ 形でない”
  if $((C_0 = C_i) \cap (S_0 = S_i))$
    $(\mathrm{Name\_part})_i =$“$(\mathrm{Name})_0$”
  else
    $(\mathrm{Name\_part})_i =$“物”
end
Reply = “$\sum_{i=1}^{n}\{(\mathrm{Color\_part})_i(\mathrm{Shape\_part})_i$
  $(\mathrm{Name\_part})_i\}$ ＋をn個見つけました．”

人に対しての質問文は以下のように作られる．

- 探索条件に合う物が1個だけ見つかった場合．
  Reply = “これですか?”
  と，人に対して確認をとる．
- 探索条件にあう物が2個以上見つかった場合．
  Reply = “どれを選びますか?”
  と，人に対して見つけた物体のうちどれを取りに行けば良いか質問をする．
- 探索条件に合うものがなかった場合．ただし，物体は1個以上は見つけている場合．
  Reply = “どうしますか?”
  これは，ロボットが見つけた物体が本当に人が望んでいる物かどうか確かめる事もしたいため，このような質問を行なう．

人からの入力があった場合の処理の例をいくつか表1にまとめる．

表 1: 質問文作成の例

		物体名	色	形	位置	数
(1)	探索物体	リンゴ	赤	球	$(\hat{x}, \hat{y})$	1
	探索結果	×	×	○	(x, y)	1
	検証結果	×	×	○	○	○
(2)	探索物体	リンゴ	赤	球	$(\hat{x}, \hat{y})$	1
	探索結果	○ ○	○ ○	○ ○	(x_1, y_1) (x_2, y_2)	2
	検証結果	○	○	○	×	×
(3)	探索物体	リンゴ	赤	球	$(\hat{x}, \hat{y})$	1
	探索結果	×	×	×	×	0
	検証結果	×	×	×	×	×

表1の case 1 の例では，シーン中に青リンゴが1個置かれている状況を考えている．この場合は，システムはリンゴは赤い物であるという知識しか持っていないため，色情報による物体認識が失敗している．(表1中には色情報による認識に失敗したことを示す(×)が書き込まれている)．したがって，色情報からは認識できなかったことを，色に関する部分を否定した文を生成することにより人に伝える．すなわち『赤色でない球形のものを一個見つけました』と返答する．すると人は対象が青リンゴなので，ロボットが色を間違っていると気がつき，「青リンゴをとって」とロボットに情報を与える返事が自然に出てくることが期待される．

case 2 の例はシーン中にリンゴらしい赤く丸いものを2個見つけた場合である．この場合は個数に関して疑問が生じている．システムはこの2つのうちどちらが欲しいか人に質問をし，人はそれに答える．

case 3 の例はリンゴらしい物が一つも見つからなかった場合である．この場合は，ロボットは赤く

て丸い物を探すために，カメラをいろいろな方向に動かす．この動作が人には「きょろきょろしているのは見つけられないからだな」という非言語的コミュニケーションになる．そこで，人はどうすればロボットがリンゴを見つけられるか考え，適切な助言をシステムに対して行なう．例えば，ロボットとリンゴの間に障害物が存在し，そのためにリンゴが見えない場合は，ロボットを別の位置にまで移動させ，そこから再度リンゴを探索することを教える．

6 実験例

提案手法の有効性を確認するために，種々の条件で実験を行なった．例として以下に示すような状況についての結果を示す．(図5参照)

- 対象物体の赤いリンゴが2個存在する．
- ロボットの方からはリンゴは重なって見える．

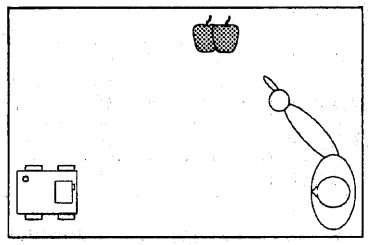

図 5: 実験例

人　　「あのリンゴをとって．」

ロボット 人の腕の方向に向く　＋　物体探索開始

{探索をした結果，リンゴは画像中では小さく写り，また重なって見えていたため，幾何学形状を特定することができなかった．}

ロボット 『赤色の球形でないものを一個見つけました．』『どうしますか?』

人　　「それをもう一度よく見て．」

ロボット ズームアップ　＋　再観測
『赤色の球形のリンゴを二個見つけました．』『どれを選びますか?』

人　　「右の物を選んで．」

ロボット 『赤色の球形のリンゴを一個見つけました．』
右のリンゴまで移動
人への確認　『これですか?』

人　　「はい．」　　<<終了>>

以上の例の他にも，対象物体が本の場合，似たような物体が複数存在するような場合など，様々な状況でもロボットはうまく対応することができた．

7 おわりに

本論文の目的は，人との対話を活用しシステムのロバスト性を向上させることである．その有効性を示すために人に代わって人に指示された物体を取りに行くロボットの開発を行なった．システムの入力には音声認識，画像認識を用いているが，特に画像認識では認識結果が曖昧になってしまう場合がある．その場合は，目標物体のところまで確実に移動できるほど認識が確定するまで，人に適当な質問を行なうことで解決する．

ただし，今回作成したシステムの仮定は大変限定されているため，今後の課題としてはその仮定を広げるために，物体認識法の改良や音声認識単語の追加などをしていく予定である．

謝辞

本研究の一部は，文部省科学研究費 (09555080, 09221219)，倉田奨励金，栢森情報科学振興財団の補助を受けた．また株式会社東芝より音声認識ソフトウェアを提供頂いた．ここに深く感謝する．

参考文献

[1] 河野恭之，屋野武秀，池田朋男，知野哲朗:「仮説推論に基づくマルチモーダル入力統合方式」，インタラクション'97 論文集，pp.33-40(1997)

[2] 伊藤敏彦，傳田明弘，中川聖一:「マルチモーダルインターフェースと協調的応答を備えた観光案内対話システムの評価」，インタラクション'97 論文集，pp.135-142(1997)

[3] 竹林洋一:「音声自由対話システム TOSBURG II －ユーザー中心のマルチモーダルインタフェースの実現に向けて－」，電子情報通信学会論文誌，VOL.J77-D-II,No.8, pp.1417-1428(1994)

[4] 金澤博史，舘森三慶，坪井宏之，竹林洋一:「雑音免疫学習を用いたサブワード HMM に基づく雑音環境下の音声認識」，日本音響学会講演論文集，pp.83-84(1996)