音声とジェスチャによる対話に基づく ヒューマンロボットインタフェース

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： 本論文はロボットに確実に作業を実行させるための，人とのマルチモーダルコミュニケーションの実現について述べる．視覚認識のような自動処理は失敗することがあるが，その場合でも人間が容易な方法で指示を与えることにより，作業が実行できる方法を提案する．例として，人に代わって物を取ってきてくれるロボットを考えた．作業中にロボットが分からないことがあれば，こ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適宜人に対して行なう．質問を作成するために，ロボットは今何が分かっていて何が分かっていないのかを把握しておくための内部状態スペースを確保している．対話には非接触型の音声とジェスチャによるインタフェースをそれぞれ組み合わせることによって，より人間同志の対話に近付けることを目指している．

Human-Robot Interface by Verbal and Nonverbal Communication

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

現在までに，ロボットに関して数多くの研究開発がなされており，特に産業用ロボットなどは今や大量生産を行なう工場などではなくてはならないものになっている．しかし，このような産業用ロボットはかなり限定された環境条件で，人によってあらかじめ指定された行動をとるだけ，というものがほとんどである．将来，今よりももっと人の指示やロボットの周りの環境の変化を柔軟に認識し，そして多種多様な作業がこなせる汎用型ロボットが求められることが予想される．このようなロボットには，多自由度の腕や制限のない移動を可能にする足回りなどのハードウェアだけでなく，外界との接点となるインタフェースの充実も必要である．本研究では，このような人にとって使い易いインタフェースの実現を目標に置いている．

環境を認識するために情報量の多い視覚情報はよく用いられている．しかし，視覚情報認識はまだまだ研究段階の部分が多く，必ずしも全て成功するとは限らない．従来のシステムでは，一回失敗してしまうと処理は完全に止まってしまい，二度と作業に復帰することができない．また，どこで失敗したのか，なぜ失敗してしまったのかも分からないことが多い．そこで，本研究では人がシステム側に適宜介入していくことを提案し，システムに与えられた処理を確実に実行させることを目標とする．

本研究で考えているロボットシステムは完全な自律型ではなく，動作中に不明な点が出てきたら，そのつど人に対して質問をすることができる対話機能を備えている．このロボットがする質問は，音声による言語的なものだけではなく，ロボットの行動(ジェスチャ)による非言語的なものによっても人に伝えられる．また人も同様にロボットに対して，音声とジェスチャを用いて自分の意志を伝える．このような複数の伝達手段によるマルチモーダルインタフェースの方が1種類の手段しかもたないインタフェースに比べてより有効であることが数々の研究で確かめられている [1][2][3].

また，この対話を実現するにはロボットは今の内部状態 (何が分かっていて，何が分かっていないのか) を完全に把握しておく必要がある．そして動作中に何か不明な点が生じた場合は，この内部状態の情報を用いて，人から最低限どのような助言が得られれば良いか考え，その助言が自然な返答の形で人から得られるような質問を生成し，実際に人に質問をする．このような対話を行なうシステムとして，東芝の音声自由対話システム TOSBURG-II[4] や，電子技術総合研究所の事情通ロボット [5] などがある．これらの研究では対話により人間の要求を機械が理解していくが，ここでは視覚情報処理の不完全さを音声とジェスチャによる対話で補うことを検討する．

本論文では，人に代わって物を取ってきてくれるロボットシステムを例にとり，人との対話方法について述べていく．人は音声とジェスチャのマルチモーダル入力によって，望んでいる物体の名前と実空間上での概略位置をロボットに伝える．ロボットは取りに行くべき物をロボット上のカメラからの画像により認識するが，この画像認識の過程で動作の不明な点が生じることが多い．

ロボット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情報の両方を用いて実空間より対象物体を探索し，正しく見つけることができた場合は，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，人に音声で質問する．また，物体が複数検出された場合も同様に，どういう物を検出したのか人に伝え，そしてどれを取りに行けば良いか質問をする．この質問に対して人は答え，システムは人の返答に応じて適切な処理を行なう．ここでは，対話の生成の方法と，実際に試作したロボットシステムによる実証について述べる．

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いるようなら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，ロボットの動作も非言語メッセージとして人に伝え

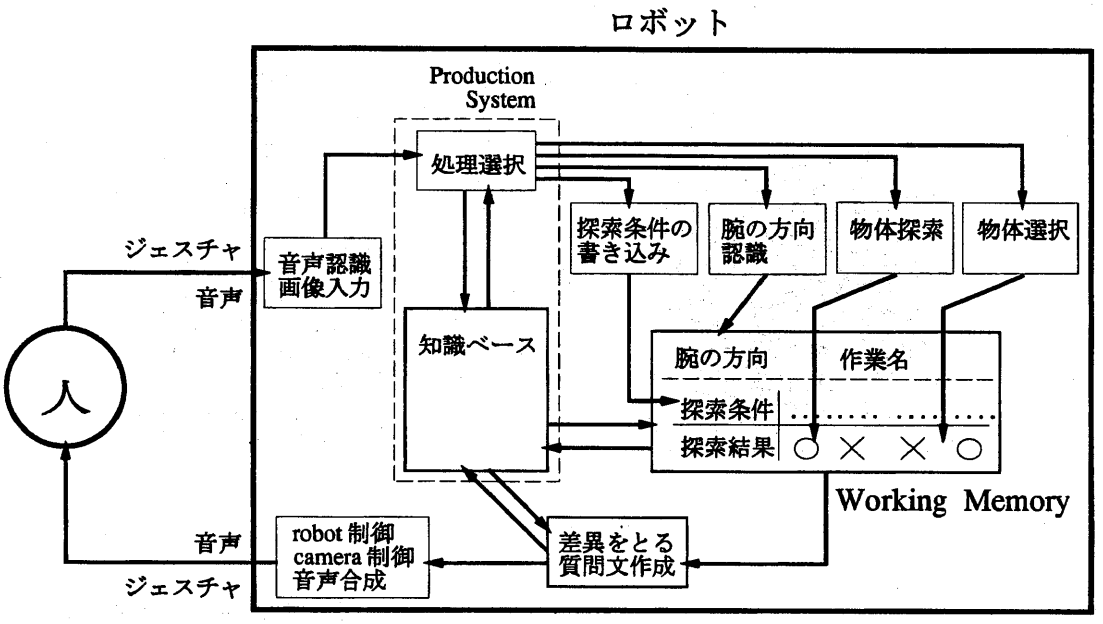

図 1: システムの概要

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，きょろきょろと辺りを見回しながら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

人の会話文は “5W1H” で構成されているが，本研究対象の指示文は主として “どこで”，“何を”，“どうする” の 3 つの要素で構成されていると考えられる．人間同志の会話で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て，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．本研究で開発したロボット内に存在する認識の曖昧性は，主にロボットの視覚部分で生じる．この原因は次のように分類される．

- 画像処理の失敗
  例えば対象物体の存在する環境が複雑で，画像中より物体を正確に切り出すことができないなど．
- 物体に関する知識の不足
  例えばロボットにとっては，リンゴは “赤くて丸いもの” 程度の知識しかもっておらず，赤色以外のリンゴは検出できない．
- 人の指示が曖昧
  複数個のリンゴを検出できたが，どれを取ってくれば良いかロボットには選択できない．

そこで，このような原因から起こる失敗を，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の 2 種類存在する [6] が，本研究も音声対話とジェスチャをコミュニケーション手段として考える．ただし，ロボットの行動も人に対してのジェスチャとして考えている．

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大きく分けて 2 つの機

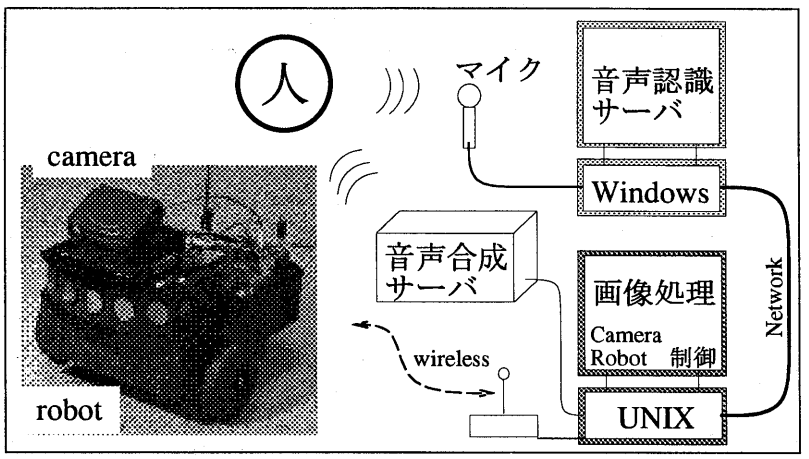

図 2: 実験システム

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. (システム入力部分)

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. (システム出力部分)

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,WM) 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ルーチン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，認識処理を開始する. 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節で詳しく述べる.

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 RS-232C 用無線モデムを搭載した. ただし，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，このロボットには物体をつかむ腕がないため，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 Windows NT 上で動作するソフトウェア [7] を使用し，その結果をネットワーク経由でワークステーション側に伝える. 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.

その他の画像認識などの主な処理はワークステーション (Sun Ultra1) で実行される.

システム内にはロボットを制御するルーチンや，音声認識を行なうルーチンなど様々なルーチンが作り込まれており，またこれらは独立に並列処理するようになっている. つまり，時間のかかる画像認識を行なっている間も常時人の言うことに対して聞

き耳を立てさせ，突然の処理の中断や処理の変更にも対応できるようにしている．以下では，これらのルーチンのうち人のジェスチャ解析と物体探索の方法について簡単に示す．

4.1 人のジェスチャ解析

今のところ人のジェスチャは，“物を指し示す”動作のみについて考えている．そのためここでは，物を指し示している腕の方向を求めることを行なっている．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，人が他人の腕の方向を見る場合も，腕の方向は完全には認識せず，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．これを考慮に入れ，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．

具体的な処理方法は，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．

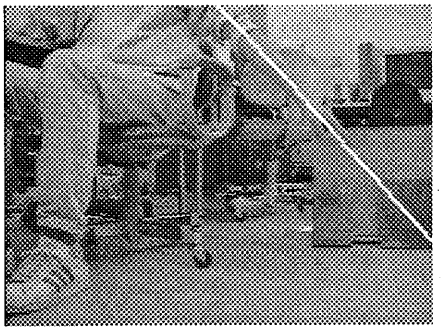

図 3: 腕の方向検出

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．

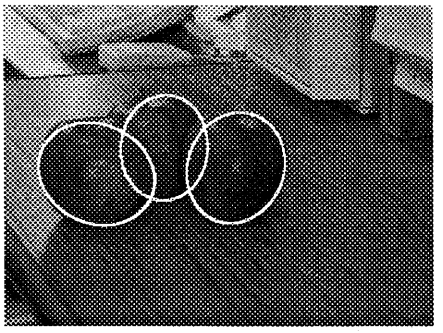

図 4: リンゴ (楕円) 検出

4.3 ロボット・カメラ制御部

ここでは，ロボットの前進，後退，回転，及びロボットに搭載されているカメラのパン，チルト，ズームを制御している．主な動作としては，人の腕の方向へ向くロボットの回転，物体の位置までの移動，物体を注視するためのカメラのズーム制御を行なう．

このロボットの行動が，人に対してのジェスチャ情報となり，人はこのロボットの動きを見て，ロボットが人の意図した通りに動作しているかどうか確かめることができる．

4.4 システム内のデータの流れ

システム内には人からの指示に関する情報や物体探索結果の情報など，様々なデータが入力される．そしてロボット制御や音声合成を行なうという形で出力される．このようなデータの処理は全て，プロダクションシステム (PS)[8] を応用して行なっている (図1参照)．つまり，例えばロボットを物体のところまで移動させる場合は，物体の位置までの行き方 (条件) を PS の短期記憶中に一度書き込み，この条件にあう処理方法 (例えばロボットを○○度回転して○○ mm 前進など) を知識ベース中より探しだして，実際にこの処理を行なうこととしている．

このように全ての情報を PS を経由して処理するのは，システムで使われる知識を一括して管理する

ことができ，また新たな機能を容易に付け加えることができるようにするためである．例えば新たに“梨”についての物体情報を付け加えたい時は，知識ベース中に

「梨は黄色で，大きさは10cm位の球形」

のような知識を加え，あと音声認識用の辞書に“ナシ”を登録するだけで済む．

また短期記憶中に，ある一定時間経過してもルールがマッチせずに情報が残ってしまうことがしばしば生じる．これは，誤った音声認識結果が送られてきてしまった場合や，突然の処理の変更など様々な原因による．このような不要な古い情報を消去できるように，それぞれの情報には短期記憶中に書き込まれた時刻に関する記述が付け加えられており，その時刻を常時監視することによって，古い情報の削除を行なっている．

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，その状態からどうすれ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．

探索条件と探索結果はフレーム形式で表現するようにしている．このフレームの構造は色，形，個数と位置の情報を要素として持つ．差異はこの探索条件に関するフレームと，探索結果に関するフレームの要素一つ一つについて比較することによって行なう．これらの差異をとり，状況説明文(Reply_a)を作るアルゴリズムを以下に示す．

for $i = 1$ **to** n **do**
  for $j = 1$ **to** m **do**
    if $(S_{0j} = \mathrm{NULL})$
      $(\mathrm{Clause})_{ij} =$“ ”
    else if $(S_{0j} = S_{ij})$
      $(\mathrm{Clause})_{ij} = (\text{affirmative clause})_{ij}$
    else
      $(\mathrm{Clause})_{ij} = (\text{negative clause})_{ij}$
  end j
  if all slots are matched
    $(\mathrm{Name\_part})_i = (\mathrm{Name\_part})_0$
  else
    $(\mathrm{Name\_part})_i =$“Object”
end i
$\mathrm{Reply\_a} =$“$\sum_{i=1}^{n}\{\sum_{j=1}^{m}(\mathrm{Clause})_{ij} + (\mathrm{Name\_part})_i +$“と”$\}$ + を n 個見つけました．”

ここで，n は見つけた物体の個数で m はフレームの構成要素の数を表す．S_{ij} は見つけた物体のうち i 番目の物の j 番目の属性を表す．また添字 i が0のものは探索条件を表す．

$(\text{affirmative clause})_{ij}$ は物体 i の属性 j を肯定する句で，$(\text{negative clause})_{ij}$ は逆に否定する句を表す．

上のアルゴリズムには表現していないが，全ての属性が条件と一致した物体は探索物体そのものであるとし，その物体を説明するための句は省略される．さらに，条件と完全に一致する物体が複数見つかった場合はそれらをまとめ，物体名とそれが何個あったかを表現するだけに簡略化し，冗長な文になることを避ける．

また，人に対しての質問文(Reply_b)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_b = “これですか?”
  と，人に対して確認をとるための文を選択する．
- 探索条件に合う物が2個以上見つかった場合．
  Reply_b = “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ための文を選択する．
- 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_b = “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ための文を選択する．

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)	2
		○	○	○	(x_2,y_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，形情報による物体認識は成功したが，色情報による認識には失敗している．(表 1中には形情報による認識の部分には成功したことを示す (○) が書き込まれ，逆に色情報による認識の部分には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『赤色でない球形の物を一個見つけました．』『どうしますか?』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物体を 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，『リンゴを二個見つけました．』『どれを選びますか．』と人に質問をし，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

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．

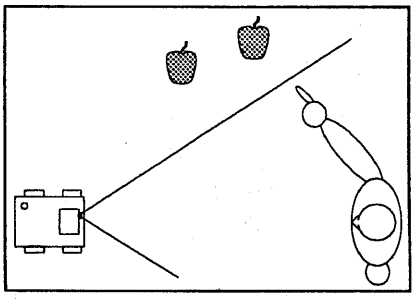

図 5: 実験例

以降，人の指示とロボットの行動についてまとめる．

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

{ 図 6は腕の方向に向いた状態を示す． }

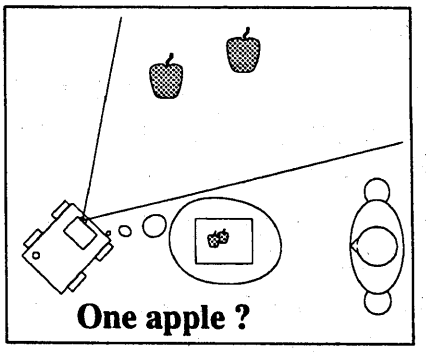

図 6: 腕の方向を向いた状態

{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
『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　『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?』
{図 7はリンゴの位置まで移動した状態を示す．}

人　「はい．」
<<終了>>

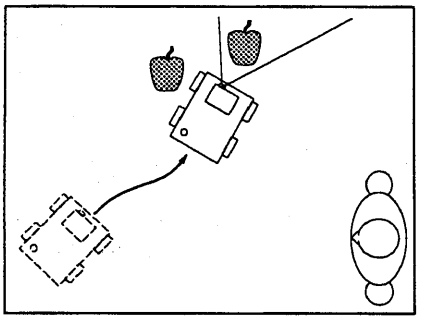

図 7: 物体位置までの移動した状態

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，特に画像認識では誤った認識結果が得られる場合がしばしば起こる．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ではまだ対象領域が限定されているため，今後の課題としてはそれを広げるために，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] 森田寿郎，渋谷恒司，菅野重樹: “人間共存型ヒューマノイド Hadaly-2 の開発 – 運動系の構築および人間との共同作業 –”，日本機械学会 [No.97-23] 第 2 回ロボティクスシンポジア講演予稿集，pp.127-132(1997)

[4] 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)

[5] H.Asoh, Y.Motomura, I.Hara, S.Akaho, S.Hayamizu, T.Matsui: “Combining Probabilistic Map and Dialog for Robust Life-long Office Navigation”, Proc. IROS96, pp.807-812(1996)

[6] 黒川隆夫:“ノンバーバルインタフェース”，オーム社，(1994)

[7] 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)

[8] 白井良明:“人工知能の理論”，コロナ社，(1992)