論　文

視覚によるサービスロボットのための簡略化発話の理解

ザリヤナ　モハマド　ハナフィア[†]　　山崎　千寿[†]　　中村　明生[†a)]
久野　義徳[†b)]

Understanding Inexplicit Utterances for Helper Robots Using Vision

Zaliyana MOHD HANAFIAH[†], Chizu YAMAZAKI[†], Akio NAKAMURA[†a)], and Yoshinori KUNO[†b)]

あらまし　生活環境内で動作するサービスロボットのヒューマンインタフェースとしては，音声対話によるものが有望である．しかし，自然で人間に負担のない音声対話システムを実現するためには，音声理解の技術だけでは十分ではない．人間同士の自然な対話では，双方の眼前にあるものについては，それに関することを言葉の上では簡略化することが多い．互いに視覚で情報を得ていることが確かな場合，それを簡略化して発話するのが普通である．また，それが許されないようでは，使いやすいインタフェースにはならない．簡略化が行われるのは，会話の当事者が行動によりかかわっており，明確にいわなくても分かると判断したからと考えられる．そこで，ものの近くにいる，ものを見ている，ものを指差している，ものを手で扱っているという行動を考え，その対象となるものを視覚情報処理で求め，それにより簡略化された発話を理解する方法を提案する．実際にロボットシステムを開発し，有効性を示す．

キーワード　対話理解，マルチモーダルインタフェース，ロボット，視線

1. まえがき

生活環境内で動作するサービスロボットのヒューマンインタフェースとしては，音声対話によるものが有望である．そこで，音声対話を用いたロボットシステムが数多く研究されている[1],[2]．しかし，まだ人間同士のような自然な対話ができるようにはなっていない．人間同士の音声言語によるコミュニケーションを考えると，常に言語で明確にすべての情報を述べているわけではない．省略や，あいまいな指示語を使うなどの簡略化した表現がよく用いられる．それでも人間同士の場合は，それでかなり意思が通じる．このような現象については，言語学や自然言語処理で多く研究されている[3],[4]．

対話における簡略化については補完処理の観点から次のような分類がされている[5]．

- 対話当事者に関する省略：叙述表現や尊敬語，謙譲語の使用に伴う省略．
- 文脈省略：補完されるべきものが，対話中の別個所で言及されている省略．
- 共有知識による省略．

（1）　対話当事者間のみにしか分からない共有知識に基づく省略．

（2）　一般的な常識から容易に推測できる事項の省略．

しかし，このような言語処理関連の分野で考えられるもの以外に，対面対話では相方の共通な五感による認識に基づく簡略化が考えられる．松尾[6]は，相手のいうことが分かるためには，手掛り情報が必要だと述べている．そして，コミュニケーション場面において，知覚できるあらゆる刺激はすべて手掛り情報になり得ると述べている．例えば，人間同士の自然な対話では，双方の眼前にあるものについては，それに関することを言葉の上では簡略化することが多い．互いに視覚で情報を得ていることが確かな場合，それを簡略化して発話するのが普通である．例えば，会話では「あれ取って」とそれ以前に「あれ」が指示する物体が言及されていないのにいうことがある．これで意思が通じるのは，「あれ」の指す物体が話し手と聞き手

[†] 埼玉大学工学部情報システム工学科，さいたま市
Department of Information and Computer Sciences, Saitama University, 255 Shimo-Okubo, Sakura-ku, Saitama-shi, 338-8570 Japan
a) E-mail: nakamura@cv.ics.saitama-u.ac.jp
b) E-mail: kuno@cv.ics.saitama-u.ac.jp

の両者に視覚で共通に認知されているからである．ここでは，このような視覚情報を共有していると発話者が考えることにより簡略化表現された発話を理解する方法を，身体の不自由な人を支援するサービスロボットの場合を例にして検討する [7], [8]．サービスロボットに限定した場合でも，上に述べたような文脈や共有知識など様々な要因による簡略化があり得る．しかし，ここでそれらすべてを考えることはできないので，そのような部分は他の言語処理研究の成果により解釈できているとして，視覚による簡略化だけが残った発話が与えられるものと仮定する．

視覚で簡略化発話を理解するための情報を得るといっても，どういう情報を視覚で得ればよいかが問題である．一般に，共同注視 [9] しているものについては，あらためてそれに言及する必要がないので，簡略化される可能性が高いと考えられる．ここでは，このような共同注視から更に広げて，話し手と聞き手が何らかの行動でかかわっているものについて，簡略化が行われる可能性があると考える．具体的には，共同注視のように見るという行為，ものを手で扱うという行為，もう少し直接的なものとして，ものを指差すという行為を考える．更に，行為を広義にとらえ，ものの近くにいるという行為も考える．以上のような行為を認識し，その行為にかかわる物体を検出することにより，その情報に基づいて簡略化された発話を理解する．

提案手法では，例えば，先に挙げた「あれ取って」の実世界での意味を理解することを音声認識と視覚情報処理を組み合わせることにより実現する．このようにマルチモーダルの情報処理で指示語を含むような発話を理解する先駆的な研究として，Bolt [10] の “Put-That-There” と呼ばれるシステムがある．これは「あれをそこに置いて」というような指示を，「あれ」といいながら対象の物体を指で差し，「そこ」というときに，今度は移動先を指で差すことにより行えるというシステムである．“Put-That-There” では，仮想世界内の対象への指示で，指差しの認識にも磁気センサを使うという点で，現実世界の対象を視覚情報処理で扱う提案手法とは違っているが，最も違う点は，“Put-That-There” では指示語に合わせて指差し動作を意識的に行うことに取り決めているという点である．提案手法でも，指差し動作の場合は，対象を指示しようとして意識的に行う行為であるが，他の行為は特に意識して対象を指示しようとしたものではない．このように自然に生じる行動の認識を利用して発話を理解しようというのが本研究の提案である．筆者らは，以前から使いやすいヒューマンインタフェースを実現するためには，意識的・意図的な行動だけでなく無意識的あるいは直接的・明示的には指示を意図したのではない行動を理解することが必要であると主張している [11]．今回の提案も，この考えにそったものである．

人間とロボットのコミュニケーションに音声と視覚を用いることについては多数の研究がある．その多くは音声認識やジェスチャ認識を用いることにより，人間が音声やジェスチャでロボットに指示ができるというものである [12]〜[14]．そこで考えられている音声・ジェスチャの多くは本研究とは違い，意識的に指示の伝達を意図した明示的なものである．例えば，McGuire ら [15] は，音声とジェスチャでロボットに把持動作を教えるシステムを提案しているが，研究の中心は音声認識とジェスチャ認識を統合して人間の指示を確実に認識しようというところにある．Roy [16] はこのようなマルチモーダルインタフェースの方向ではなく，対象物をロボットに見せながら，そのものの名前をいうことにより，音声と実世界の物体の対応をとり，対話を通じて実世界を理解できるようになっていくロボットを提案している．これは非常に興味深い研究であるが，本研究で考えているような人間の依頼発話の理解は考えていない．

本論文では，以上のような視覚によるサービスロボットのための簡略化発話の理解法について述べる．そして，実験により有効性を確認する．

2. 簡略化発話

本研究で扱う簡略化発話は，省略と直示の二つに分けられる．省略とは，文中の主語や述語が省かれる現象をいう．つまり，言語を運用するにあたり，ある部分を必要としなくとも言語の本来の機能である情報の伝達を行うことができることである．話し手の意図することが表現形式としては完璧ではないのにもかかわらず聞き手は何らかの理由，過程を経て理解することができるとき「省略」という現象が現れてくる．視覚による省略の一例を図 1 に示す．この例では，ユーザはいきなり「4 にして」という．この文だけからでは何を意味するのか分からない．しかし，これがテレビの方を見ながら発話されたとすると，テレビのチャネルを 4 にしてという意味と推測することができる．

直示とは一般にジェスチャ（指，視線など）で現実（実世界）の対象を指示することである．つまり，言葉

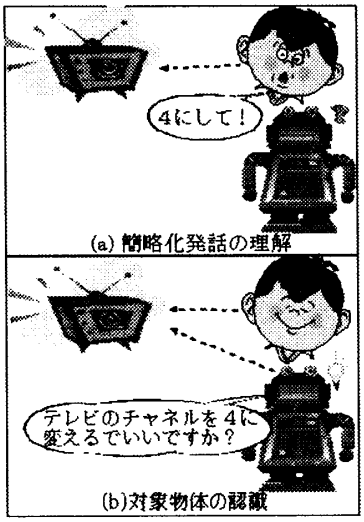


図 1　ケース 4 の簡略化発話の例
Fig. 1　Inexplicit utterance example (Case 4).

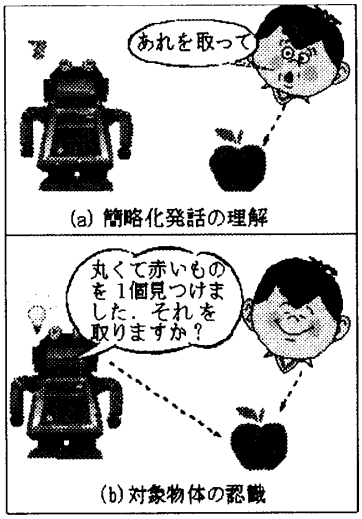


図 2　ケース 8 の簡略化発話の例
Fig. 2　Inexplicit utterance example (Case 8).

と世界をつなぐ働きをする．人は直示的な発話をするとき，発話を簡略化することが考えられる．日本語の直示詞（代名詞）は 3 種類（コ系・ソ系・ア系）に分けられる．本論文では「代名詞直示：deictic pronouns」と「空間的直示：spatial deictic」を扱う．この二つの直示は実世界で存在するものを指示する言語照応である．例えば，指差しながら「あれを取って」と依頼する場合，「あれ」というのは代名詞直示であり，実世界に存在するものを指示するときに使用する．空間的直示の場合は，「テーブルの上にあるものを取って．一番上」のように空間的な位置の発話をして，対象物体の位置を照応する直示である．直示で対象物体を示す場合の例を図 2 に示す．

表 1　発話の分類
Table 1　Utterance patterns.

⟨対象⟩, ⟨動詞⟩	不明 ⟨×⟩	あいまい ⟨△⟩	明確 ⟨○⟩
不明 ⟨×⟩	1. ⟨×, ×⟩	4. ⟨×, △⟩	7. ⟨×, ○⟩
あいまい ⟨△⟩	2. ⟨△, ×⟩	5. ⟨△, △⟩	8. ⟨△, ○⟩
明確 ⟨○⟩	3. ⟨○, ×⟩	6. ⟨○, △⟩	9. ⟨○, ○⟩

ここでは，サービスロボットに依頼をすることを考えているので，ロボットに伝えたい情報は依頼の内容となる行為を示す動詞とその動詞に関する対象物（目的語）になる．発話中に，この両者がどのように表現されているかを場合に分けると表 1 のようになる．動詞と対象のそれぞれについて，言葉で明確に表現されている場合，言葉が省略されている場合，更に，ある程度言及されているが，それだけでは明確には分からないあいまいな場合がある．ここで考えているあいまいな場合は，対象については，上述の直示である．また，動詞については，「する」「やる」など，それだけいわれても依頼内容がはっきりと分からないものである．

表 1 のそれぞれの場合の例を以下に挙げる．

C1. 対象不明；動詞不明 ⟨×, ×⟩.
例：“こんにちは”

C2. 対象あいまい；動詞不明 ⟨△, ×⟩.
例：“あれ”

C3. 対象明確；動詞不明 ⟨○, ×⟩.
例：“そこのリンゴ”

C4. 対象不明；動詞あいまい ⟨×, △⟩.
例：“4 にして”（図 1 参照）

C5. 対象あいまい；動詞あいまい ⟨△, △⟩.
例：“あれをやって”

C6. 対象明確；動詞あいまい ⟨○, △⟩.
例：“赤いのにして”

C7. 対象不明；動詞明確 ⟨×, ○⟩.
例：“持ってきて”

C8. 対象あいまい；動詞明確 ⟨△, ○⟩.
例：“あれを取って”（図 2 参照）

C9. 対象明確；動詞明確 ⟨○, ○⟩.
例：“あの赤い本を取って”

3. 視覚による情報の獲得

2. で述べたような簡略化された発話がなされた場合，そのような発話でも視覚情報により分かるはずと人間が考えたからだとみなす．他の要因による簡略化は，自然言語処理などの研究成果により解決されるも

のとし，ロボットへの発話には視覚に起因するものだけが残されていると仮定する．

はじめに対象の部分の簡略化について考える．対象の部分についての簡略化はあいまい（直示）と不明（省略）の二つに分けてはいるが，今回のサービスロボットへの依頼という領域では，ロボットにとっては，どちらも簡略化された表現の中の対象と実世界の物体を対応づけなければいけないという点では同じである．ここでは日本語を考えているが，日本語では，例えば，「あれ取って」の代わりに単に「取って」ということもできる．「あれ」という言葉自体は省略されているが，依頼発話の理解という点では，直示と同様と考えられる．したがって，以下では両者をまとめて簡略化として考えることにする．

対象の部分は，その環境中にある物体のはずである．それを簡略化したということは，人間またはロボットが何らかの行為でかかわっているので，それについていう必要がなかったと考えられる．人間またはロボットの行為と，それにより特定される対象としては，以下のものが考えられる．

V1. 近くにいる.（これは一般的な言葉の用法では行為に入らないかもしれないが，人間あるいはロボットが物体の近くにいるという行為である.）人間またはロボットのすぐ近くにあり，それについて言及していることが明らかであると考えられる物体.

V2. 見ている. 人間が見ている物体（人間がロボットもその物体を見ていると思っている場合）．視線や顔の向きで見ている物体は推察できる.

V3. 指差している. これは意識的に行動をしている場合で，当然，対象は指で差された物体である.

V4. 操作している. 人間またはロボットが単に触っているなども含めて何らかの操作をしている物体.

発話された音声依頼文中に対象物体に関する情報（実世界の情報）が足りなかったと判断された場合，上記で挙げた人の行動にかかわる物体を検出する．異なる行動に対して複数の物体が検出されたときには，指差し動作があれば，それは指示を意図した動作なので，それに関する物体が最優先される．簡略化された依頼発話に日本語の代名詞のコ系・ソ系・ア系が含まれている場合は，表 2 に示す優先順位で対象物体の候補とする．コ系は話し手の近く，ソ系は聞き手の近く，ア系は両者から離れたものを指すのが普通だからである．これ以外の場合については，複数の候補物体が検出されたときは，現在は，音声で人間に聞くようにしている．この点は今後の検討が必要である.

表 2　直示詞と行動の関連
Table 2　Relation of pronouns (deixis) and human actions.

分類	行動
コ系（例：これ）	1）ユーザが操作している物体 2）ユーザの近くにある物体 3）視線
ソ系（例：それ）	1）指差し 2）ロボットの近くにある物体 3）視線
ア系（例：あれ）	1）指差し 2）視線

表 3　カテゴリーと登録単語の例
Table 3　Dictionary.

カテゴリー	登録単語
命令（動詞）	明確：取って，持ってきて，置いてきて あいまい：して，やって
遊離対象（名詞）	物体名：本，リンゴ，リモコン 属性：赤い，黄色い，円形，四角形
付着対象（名詞）	本棚，テーブル，テレビ，冷蔵庫
その他	あいさつ（感動詞），数字（名詞-数）など

表 3 に示すように辞書に登録した物体（名詞）は，場所に固着して動かない物体（付着対象：本棚，テレビなど）と，簡単に動かすことができる物体（遊離対象：本，リンゴ）に分けておく．アフォーダンスの提唱者の Gibson は，視覚で知覚されるものは，色，形態，位置，空間，時間，運動などではなく，場所（place），付着対象（attached object），遊離対象（detached object），持続する物質（persisting substance），事象（event）であると主張している [17]．ここでは，筆者らのグループの以前の研究と同様にこの Gibson の分類を利用する [18], [19]．その研究では，遊離対象は視覚で認識したが，付着対象は一般に大きくて小さな単一領域として画像処理で検出しにくいため，視覚では認識せずに，ユーザにその位置を音声で教えてもらっていた．ただし，付着対象は動かないので，一度教えてもらえば，その位置を地図に書き込むことにより，2 回目以降はユーザにその位置を教えてもらう必要はなくなる．今回の研究は，その研究の発展なので，付着対象については，既に位置が分かっているものとする．**5.** で述べる視覚情報処理で，視線方向や指差し方向に付着対象が見つけられない場合は，ロボットは地図を用いて，その方向に付着対象がないか調べる．もし，地図に当該物体が記載されていれば，それを視線や指差しの行為にかかわる物体と考える.

動詞の部分は，対象が決まれば，それに対してできる行動ということで可能なものは限定されてくる．したがって，対象物体ごとに可能な行動を辞書に登録しておき，それを推定値として用いる．ただし，人間が対象物名を言葉でいった場合や，上で述べた付着対象の場合は物体名が分かるが，その他の場合は物体名は分からない．すなわち，対象が簡略化されて発話された場合，ロボットは行為にかかわる物体を検出することにより対象は分かるが，それが何かまでは分からない．その場合，デフォルト値として，遊離対象については「取ってくる」を用いることにしているので，これが動詞の候補になる．詳しくは **8.** で述べるが，提案のシステムは，行動を起こす前に音声で確認をとることにしており，また，分からないときにも音声で聞くようにしている．付着対象は物体名が分からないことはないが，対応する動詞のデフォルト値は「(そこに) 行く」としている．なお，辞書に書かれた情報だけからでは一つの動詞にしぼれない場合は，人間に音声で聞くことにしている．

4. 音声対話システム

対話によるインタフェースを実現するため，IBM Via Voice SDK for Windows が，Visual Basic 上で音声入出力を利用するために提供している ActiveX コントロールを用いて音声処理部を作成した．これを利用することにより，マイクから入力された音声は文字列に変換されて以降の処理に利用される．

文字列に変換された音声発話は形態素解析処理を行い，そして **2.** で説明した発話の可能性の分類（9 ケース）に基づき，依頼文のパターンを判断する．形態素解析部分は奈良先端科学技術大学院大学で開発された茶筌を利用した [20]．

図 3 で完全な依頼文は {<対象 (○)>, <動詞 (○)>} と表される．形態素解析の結果と登録した単語の情報を用いて，表 1 に示した発話パターンに分類する．例えば，図 3 中の例 2「4 にして」という発話の場合，形態素解析の結果は {<その他の品詞–名詞–数>, <動詞>} になる．そして，「して」という動詞が表 3 であいまいな動詞として登録されているので，発話の分類は C4 {<対象不明 (×), <動詞あいまい (△)>} になる．

5. 視覚情報処理

言語処理の結果，対象が明確でない場合は，**3.** で述べた V1～V4 の視覚情報処理が起動される．ここでは，図 4 のように移動ロボット（ActivMEDIA 社 Pioneer2）に 2 組のステレオカメラを搭載したものを用いた．

2 組のステレオカメラの使用法を図 5 に示す．下段

記号の説明は以下の通りである．
S → *Sentence*: 文
V → *Verb*: 動詞
N → *Noun*: 名詞–一般
Adj → *Adjective*: 形容詞
P → *Noun-Pronoun*: 名詞–代名詞
O → *Others*: その他の品詞（感動詞，名詞–数）等
NP → *Noun Phrase*: 名詞句

■**属性**
Adj → [形容詞–自立]
Adj2 → [名詞–一般] + [形容詞–接尾]
■**対象物体名**
N → [名詞–一般]
■**指示**
V → [動詞–自立]
else V → [動詞–自立]+[助詞–接続助詞]
else V → [動詞–自立]+[助詞–接続助詞]+[動詞–非自立]
■**物体の属性**
NP → *Adj N*
else NP → *Adj2 N*
■**依頼文**
完全な依頼 { ⟨ 対象 ⟩ ⟨ 動詞 ⟩ }
例 1： 赤い本を取ってきて { ○, ○ }
S1 → *adj N V*
例 2： 4 にして { ×, △ }
S2 → *O V*
例 3： あれ取って { △, ○ }
S3 → *P V*
例 4： そこのリンゴ { ○, × }
S4 → *P N*

図 3 簡略化発話の判断
Fig. 3 Classification of inexplicit utterance cases.

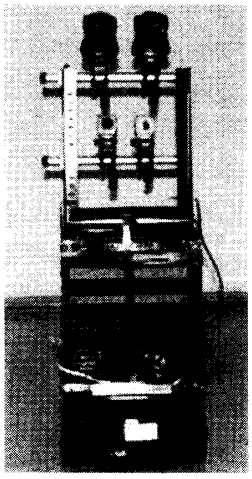


図 4 ロボットの外見
Fig. 4 Robot system.

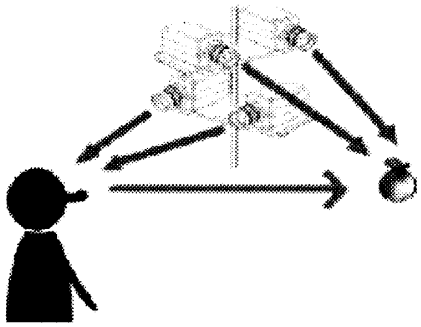


図 5　2 組のステレオカメラの使用法
Fig. 5　Use of the two pairs of stereo cameras.

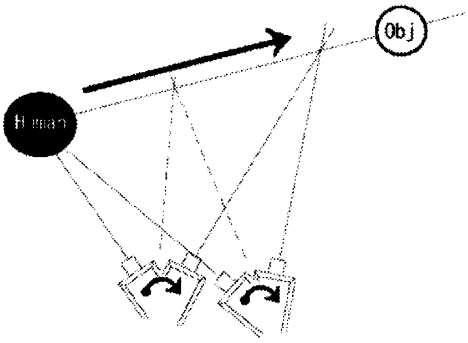


図 6　視線方向にある物体の検出
Fig. 6　Detection of an object in the gaze direction.

図 7　実験シーン
Fig. 7　Experimental scene.

図 8　物体の認識結果（左：ステレオ左画像，中：ステレオ右画像，右：ZDF 出力，中央の四角が検出された物体）
Fig. 8　Object detection result. (Left: Left stereo image, Center: Right stereo image, Right: ZDF output, The square shows the detected object.)

の一組（IEEE 1394 カメラ ソニー DFW-V500）は常に人間の方に向けられ，顔の向き（概略の視線方向と考える）や手の動きを求めるのに使われる．上段の一組は，2 台のパンチルトカメラ（ソニー EVI-D100）からなり，人間以外の環境を見る．主に，下段のカメラから得られた顔の向きや指差しの向きに存在する物体を検出する．

下段のステレオカメラでは，九州大学で開発された MARIO システム [21] を利用し，それにプログラムを追加することにより，指差しジェスチャの認識と顔の向き（概略の視線情報）を求められるようにしている．また，手を追跡し，手の近くにある物体を検出する．これが V4 の対象になる．これには現時点では [18], [19] で用いた色と形に基づく簡単な検出処理を用いている．

上段のステレオカメラでは，下段のカメラから顔の向き，あるいは指差しの向きが求められたら，三次元世界でその向きが示す半直線を考え，図 6 に示すように，ステレオの二つのカメラを光軸がその線上で交差するように，人間に近い方から遠方に向けて回転させる．そして，その際に得られる左右画像の中心部の相関を調べ，相関値が大きければ，物体があるとする．すなわち，指差しや視線の先を zero-disparity filter (ZDF) [22] で調べて，その方向の物体を検出する．これが V2 あるいは V3 の対称物体の検出になる．上段のカメラでは，更に人間あるいはロボットの近くの物体を V4 の対象検出に用いたのと同様の方法で求め，V1 の対象物体検出としている．

図 7 に実験シーンの一例を示す．人間の顔の向きを ZDF により調べることにより，図 8 に示すように電気ポットが検出されている．

6. 視覚と音声の同期

5. で V1〜V4 の物体検出法を述べたが，行動のいつの時点でかかわった物体を発話理解に利用するか検討する必要がある．基本的には，下段のステレオカメラでは常に人間の顔と手を追跡しておき，他の処理はその情報を用いて簡略化発話が発話されたと判定されたときに行う．

V1，V4 の対象物体は発話の最中と直後で変わることはあまりないと考えられるので，簡略化発話と判定された後，その時点の手や人間及びロボットの位置情報をもとに処理を行う．

V3については，指差し動作は発話に伴って行われるのが普通なので，発話の際に手が挙げられる動作を指差しジェスチャと認識する．そして，その指先の方向を求め，それからZDFでその線上の物体の検出処理を行う．

V2の場合，視線や顔の向きはV1，V4の身体や手に比べて速く動くので，どの時点の方向にある物体を対象物体とするかが問題になる．顔の向きは絶えず求めているが，顔の向きや視線は発話の最中にも変化する．簡略化発話を理解するために，どの時点の視線方向を調べればよいのか決定する必要がある．この決定法を考えるために，以下のような実験を行った．

シーン中に五つの物体を置き，被験者にロボットに対して「あの（物体名）取って」と，その物体を取ってもらいたいと想定していってもらった．ロボットは実際にその物体を取ってくることはしないが，発話の前から発話の後まで，被験者の顔の向きを求めるようにしておく．被験者は筆者の学科の学生3名で，五つの物体から毎回ランダムに選んだ物体について依頼してもらうことを，各被験者につき20回行った．すなわち，総計60回の場合について，発話と視線（顔の向き）の関係を調べた．

図9に実験の結果を示す．この図では10フレーム（0.33秒）ごとに時間を区切り，その中で主として見た方向の割合（60回の場合に対する百分率）を示している．特定の方向に5フレーム（0.17s）未満の時間しか滞留しない場合は，視線が移動中（図9ではMovingと表記）とした．移動中以外の視線方向としては，対象となる物体，依頼の相手であるロボット，それ以外の三つの方向に分けた．図で分かるように，人間が見るのは主に対象物体と対話の相手のロボット

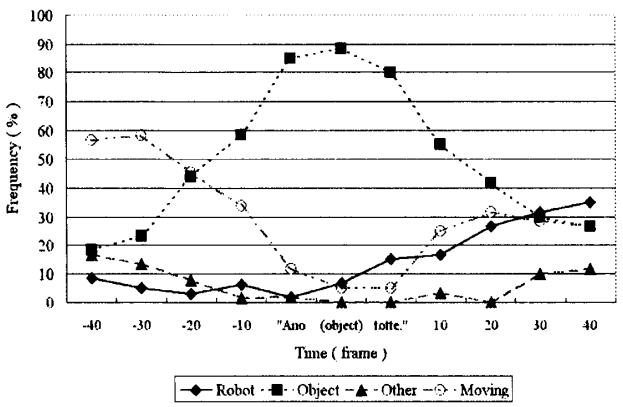


図9　発話中の視線方向の変化
Fig. 9　Gaze direction changes when giving an order.

である．発話の中で対象物体名をいう時点では90%程度の割合で物体の方を見ている．注目されるのは，対象物体の方を発話が始まる前に見始めることが多いことである．これはKaurら[23]の視線と音声による入力システムによる実験と符合する結果である．

簡略化発話の際の視線を調べるためには，例えば「ちょっと取って」（C7），「あれ取って」（C8）というような発話の際の視線方向を調べるのが望ましいが，被験者にある物体を取ってきてもらうことを念頭において，このような発話をしてもらおうとすると，対象物を意識しすぎてしまい，その物体の方だけ見てしまうことが観察された．そこで，今回は手始めとして上述のような実験を行った．簡略化発話ではないが，この実験から，発話の少し前の時点から発話中の視線方向を求めれば，その間の主な視線方向のうち，ロボットの方でないものが，対象物体の方を指していると考えられる．そこで，現時点では，簡略化発話が終わった時点で，発話の始まる少し前からの視線方向からロボット以外の主な方向を求め，その方向の物体をZDFで検出することにした．

7. 簡略発話理解の流れ

視覚情報処理を用いた簡略発話の理解について，これまで部分ごとに述べてきたので，ここで全体をまとめる．図10に処理の流れを示す．

音声認識結果を解析し，発話をC1～C9のパターンに分類する．対象も動詞もないC1の場合は，あいさつなどが考えられる．この場合は認識された単語（列）を辞書で検索し，もし辞書にあれば，そこに書かれた定型的な応答をユーザに返す．例えば，あいさつなら，あいさつの返答をする．もし，辞書になければ，「何ですか，何か御用ですか」と返答する．

C3，C6，C9のように対象が明確にいわれている場合は，まずその対象が付着対象か辞書で調べる．もし，付着対象なら地図にそれが載っているか調べる．載っていない場合は，以前に発表したシステム[18],[19]と同じく（実際には，そのシステムに今回の簡略化発話理解の部分を追加して，現在のシステムになっている），ユーザに対話を通じてその場所を教えてもらう．付着対象でない遊離対象の場合は，その物体を検出する視覚情報処理を行う．この場合も，認識できない場合は，以前のシステムと同様にユーザとの対話により，物体を検出する．以上で発話の中の対象が実世界の物体と対応づけられたら，次に動詞が不明（C3）あいま

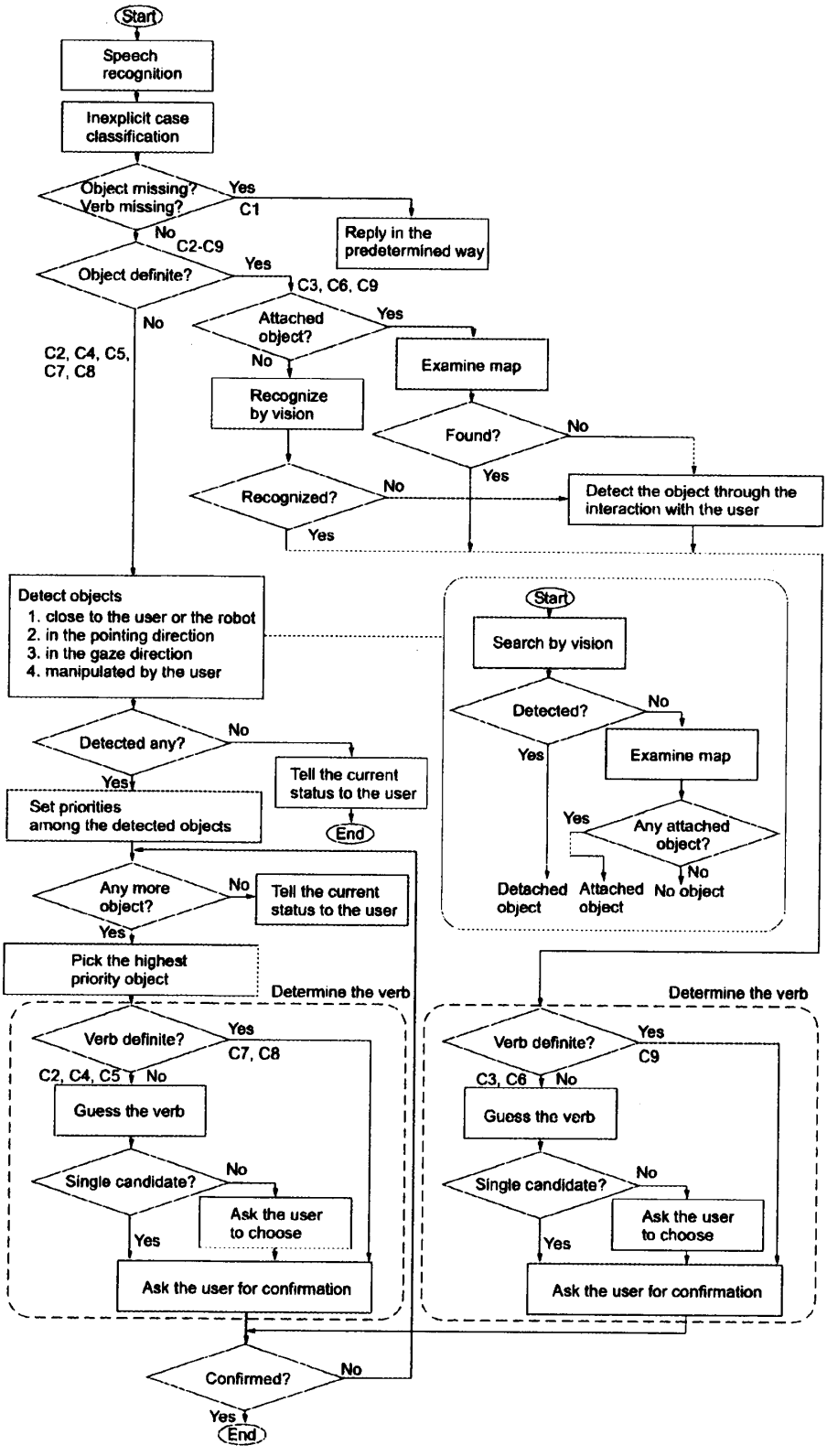


図 10　視覚情報処理を用いた簡略発話理解の流れ
Fig. 10　Flow of understanding inexplicit utterances using vision.

い（C6）の場合は，動詞を推定する．これは，辞書に対象ごとに可能な動詞が挙げられているので，それを推定値とする．もし，複数の可能な動詞が辞書に書かれていて，それから選択できない場合は，辞書の動詞を読み上げ，どれにするかユーザに聞く．選択できる場合というのは，発話中に他の情報があり，その情報

と動詞の関係が辞書に書かれている場合である．例えば，テレビの場合なら，「スイッチをつける，消す」，「チャネルを変える（数）」，「音量を変える（大きく，小さく）」などの可能な動詞（動作）がある．ここに，括弧内に記したように関連する語が挙げられている．それにより，例えば，「テレビを 4 にして.」（C6）という発話は「テレビのチャネルを 4 に変えてほしい.」ということだと理解する．現在の実装では，どんな場合も，最後にロボットが理解したことをユーザに告げて，それでよいか確認をとるようにしている．

次に，対象があいまい（C2，C5，C8），不明（C4，C7）の場合について述べる．これらの場合は，近隣（V1），視線（V2），指差し（V3），操作（V4）に関連する物体を視覚情報処理で検出する．視線と指差しに関しては，その方向に物体が見つからなかった場合は，地図にその方向に付着対象が記録されていないか調べる．もし，付着対象があれば，それを検出された物体とする．もし，一つも物体が検出できないときは，依頼が分からないことをユーザに告げる．物体が検出された場合は，検出された物体の間に優先順位を付ける．指差し方向に検出された物体があれば，これが最優先される．C4，C7 の場合は，その他の物体間に順序を付ける強い根拠はないが，一応，現在の実装では V4，V1，V2 の順にしてある．C2，C5，C8 の場合には，表 2 のように直示詞のコ系，ソ系，ア系に応じて順位を決めている．優先順位が決まったら，その順で，それぞれの対象ごとに以後の処理を行う．まず，動詞があいまい（C4，C5），不明（C2）の場合には動詞の推定を行う．この部分は対象が明確な場合（C3，C6，C9）で述べたものと同じである．これにより，例えば，テレビの方を見ながら，「4 にして」（C4）といえば，「テレビのチャネルを 4 に変えてほしい」というように理解できる．最後に，理解の結果をユーザに告げて確認をとる．もし，違っていれば，次の優先順位の物体について同様の処理を繰り返す．

8. 実　　験

8.1 簡略化発話の理解

5. で述べたロボットを用いて，簡略化発話の理解の実験を行った．ロボットで実際に依頼された作業を実行するには，作業に必要なマニピュレーション等に関して別の多くの研究開発が必要なので，今回は視覚情報処理により簡略化発話を理解できれば成功と考えた．

2. で述べた C2 から C8 の様々な場合について実験を行い，四つの行動に関する物体を認識することで簡略化された依頼発話を理解できることを確認した．以下に，四つの行動に関しての結果の例を示す．ただし，C3，C6 の対象が明確な場合は，対象を実世界から見つける部分は以前のシステム [18], [19] で扱った問題であり，動詞の推定の部分には視覚情報処理が関連しないので，結果の記述は省略する．また，対象と動詞のともにあいまいな C5 は，代表的な例として，「それをして」，「あれやって」というようなものが考えられるが，これらは会話の文脈があって使われ，そしてその中で理解されるもので，今回検討した視覚にかかわるような例があまり考えられない．したがって，これについては形式的に処理できることは確認したが，実際のシーンでの実験例はない．

1. 物体の近くにいる　「それこっちに持ってきて」（C8）

図 11 に示す状況で，ユーザが「それこっちに持ってきて」といった場合．この例では，「それ」という代名詞があるのでロボットの近くの物体ということで，図 12 に示すように赤いファイルが視覚情報処理により検出されたので，それを持ってきてほしいことだろうと依頼を理解した．現時点のロボットではファイルということは分からないので，「赤い色の四角形ものを 1 個見つけました．これを取りますか」とユーザに音声で確認した．以後の例でも，遊離対象については物体名は認識できないので，このように検出した物体の色と形によりユーザに確認を求め，その内容が正しければ成功とした．

2. 見ている（視線）　「あれ取って」（C8）

6. の図 7 の状況で，ユーザが「あれ取って」といった場合．視線方向を ZDF で調べ，図 8 に示す電気

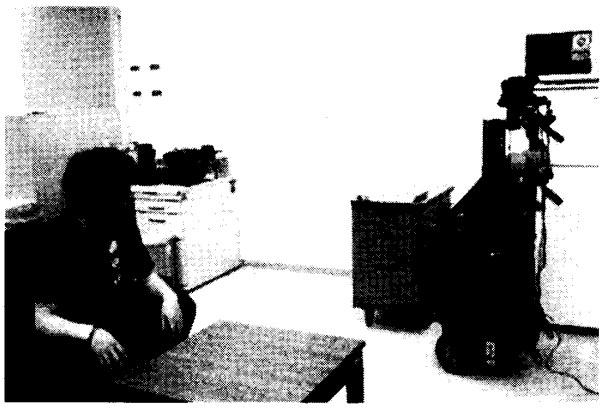


図 11　ロボットの近くの物体
Fig. 11　Object close to the robot.

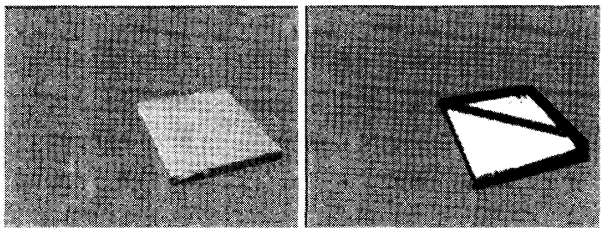


図 12　画像処理結果
Fig. 12　Detected object.

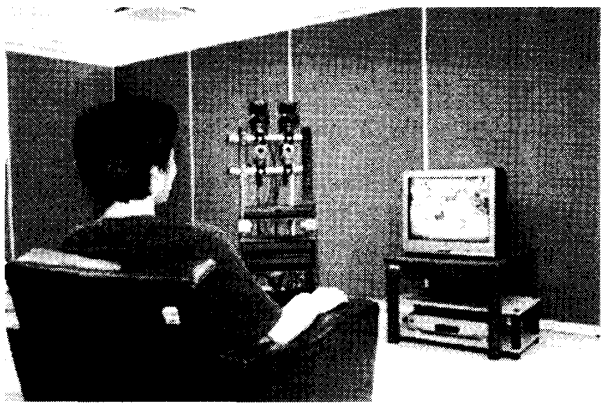


図 13　テレビ聴取シーン
Fig. 13　TV-watching scence.

ポットを検出し，それを取ってきてほしいものと理解できた．

2′. 見ている（視線）　「4 にして」(C4)

図 13 に示すようにテレビを見ているときに「4 にして」といった場合．この場合，視覚情報処理では遊離対象は検出できなかった．しかし，視線方向について地図を調べるとテレビがあり，辞書のテレビの項目の内容から動詞を推定し，テレビのチャネルを 4 に変えてほしいということだと理解できた．

3. 指差している　「取って」(C7)，「あれ」(C2)

図 14，図 15 に示すように，ユーザが指差しながら「取って」といった場合．図 16 に示すように指差し方向にあるトマトを検出して，これを取ってくることだと理解できた．また，同様の状況で「あれ」といった場合，指差し方向に遊離対象である赤い丸い物体を検出したので，遊離対象に対するデフォルトの動詞の推定値の「取る」を採用し，ユーザに「赤い色の丸いものを 1 個見つけました．これを取りますか」と確認できた．

4. 操作している　「これ捨てて」(C8)

図 17，図 18 に示すようにユーザが缶を持ちながら，「これ捨てて」といった場合．図 19 に示すように，ユーザが持っている缶を視覚情報処理で検出し，それ

図 14　指差し行動シーン
Fig. 14　Pointing scene.

図 15　指差し行動
Fig. 15　Pointing action.

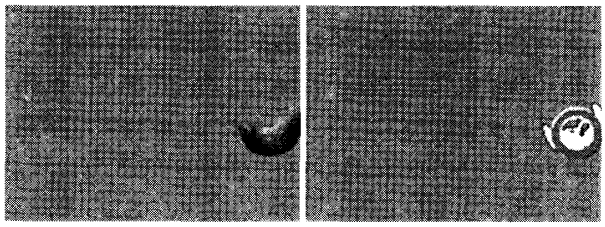


図 16　画像処理結果
Fig. 16　Detected object.

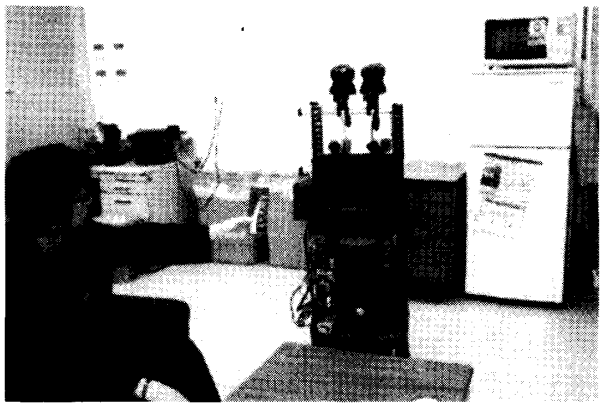


図 17　操作行動シーン
Fig. 17　Manipulation scence.

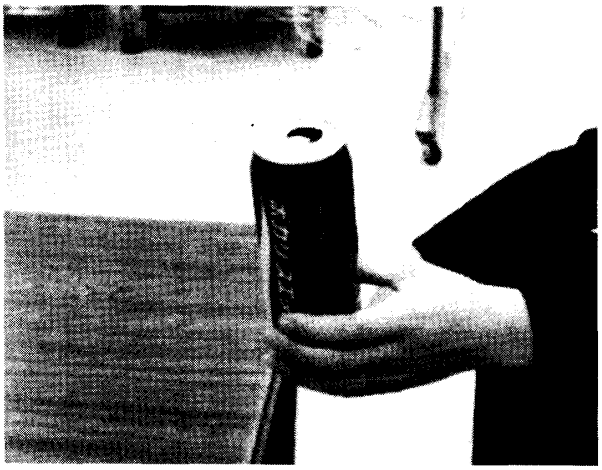


図 18　操作行動
Fig. 18　Manipulation action.

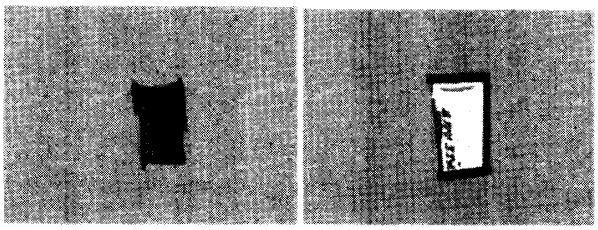


図 19　画像処理結果
Fig. 19　Detected object.

対象物体を見ながら，発話している．
ユーザ：あれを取って．
ユーザの視線方向にある物体を検出して，その処理の結果をユーザに知らせる．検出された物体は目的対象物体かどうかをユーザに確認する．
ロボット：赤い色の四角形のものを 1 個見つけました．これを取りますか?
ユーザ：はい．
ロボット：了解しました．

図 20　視線を用いた補完方法による会話結果
Fig. 20　Dialog example by the proposed method.

を捨てることだと理解できた．

8.2　ユーザの負担の軽減

簡略化発話が使えるとどれくらいユーザの負担が軽減できるか確認するための実験を行った．例として視線方向に赤色のファイルを置き，それを取ってきてもらう場合について，提案の方法と，以前に開発した対話により対象に関する情報を集めて依頼を理解するロボット [18], [19] の場合で，どういう対話になるか調べた．前者の場合を図 20 に，後者の場合を図 21 に示す．ここでは推定が正しかった場合を調べたので，当然ながら提案の方法の方がより短い対話で依頼が伝わり，ユーザへの負担が小さくなっている．同種の対話

ユーザ：あれを取って．
ロボット：物体を教えてください．
ユーザ：あのファイル
ロボット：ファイルの場所と特徴を教えてください．
ユーザ：テーブルの上．
ロボット：ファイルの特徴を教えてください．
ユーザ：テーブルの上の赤いファイル．
ロボット：テーブルの上の赤いファイルを取るですね．
ユーザ：はい．
ロボット：了解しました．
画像処理を起動し，色情報を元に物体検出を行う．検出結果をユーザに述べ，目的対象物かどうかユーザに確認する．
ロボット：赤い色のファイルを見つけました．これですか?
ユーザ：はい．
ロボット：了解しました．

図 21　人に問い返す方法による会話結果
Fig. 21　Dialog example by the conventional method.

を 10 回行った結果では，提案方法での対話の平均時間が 17.3 秒で済んだのに対し，従来法では平均 44.6 秒かかった．

8.3　現状と課題

現時点の研究は，300 単語程度の登録された簡単なシステムを開発し，上記のような実験により，行動にかかわる物体を視覚情報処理で認識することにより，簡略化発話が理解できる見込みであることを確認した段階である．今回はサービスロボットのヒューマンインタフェースと限定したので，人間の発話が依頼を表すものと制限できた．したがって，発話を九つのパターンに分けることで，共有視覚情報による簡略化発話を理解することができた．更に一般的な発話を理解するためには，基本的には対話の参加者が何らかの行動でかかわることにより共通に知覚されている情報に関して簡略化が行われるという本研究での提案を進展させたものになるが，発話内容と理解に必要な視覚情報の関係について検討を深めなければならない．また，今回の研究では簡略化発話の理解法に重点を置いたので，視覚情報処理は必要な情報を得られるものと仮定している．実際には，背景に対して顕著なほぼ単色に近い色を有し，形状も簡単な物体を実験の対象物にすることで，**5.** に述べたように既存の処理法やそれに若干の改良を加えた方法で物体が検出できるようにした．また，顔や手を追跡したり顔の向きを求めるのにも同様に既存の手法をもとにした（したがって，今回の論文では視覚情報処理の部分は簡単にふれるにとどめた）．実際に有効なシステムを実現するためには，この視覚情報処理の部分を強化する必要がある．

提案の方法は筆者らのグループが以前に発表した，

ユーザとの対話を通じて依頼を理解していく枠組みの中に組み込まれている[18],[19]．それは，更に筆者らの初期の研究をもとにしている[24]．これらの研究では，複雑な環境では失敗することが多い視覚情報処理に対して，ユーザとのインタラクションを通じて成功に導くことを検討した（言語だけでなく指差しなどの非言語行動も使うので，対話でなくインタラクションという言葉を使うことにする）．このように，筆者らは基本的には実環境で動作するシステムを実現するためにはユーザとの対話，インタラクションが不可欠であると考えている．しかし，ユーザにあまりにも多くのことを聞かなければならないのでは，使いやすいシステムとはいえない．今回の提案でも，複数の物体が検出されたり，複数の動詞候補がある場合は，ユーザとの対話でしぼるようにしている．前にこれらの解決は将来の研究課題と述べたが，これは，このようなインタラクションは本質的に必要だが，ロボットの方で更に限定できたり，優先順位が付けられれる方が使いやすいシステムになるという立場からの記述である．実際に，対象か動詞のどちらかが明確に分かっていれば，他方に複数候補があっても，優先順序付けが可能な場合があると思われる．これについては今後，研究を進めていきたい．

ユーザとのインタラクションを考える際には現実の物体と名称の関連付けについても検討する必要がある．現時点では対象についてはユーザが明確にいわなければ，もの自体は視覚情報処理で検出できても，その名称はロボットには分からない．したがって，現在の実装では，ロボットは対象について言及するのに，その属性を列挙している．例えば，ユーザがリンゴを指差して，「それ取って」といったとする．ロボットは指差しの先に赤い丸いものを見つけると，「赤い色の丸いものを1個見つけました．これを取りますか」とユーザに確認することになる．これで会話が終了すれば，大きな問題はないが，それ以降にもその物体についての話が続くようなときに，ロボットがいつもその物体を「赤い色の丸いもの」というのでは，人間にとっては耳障りである．例えば，先のロボットの確認の後に，ユーザが，「そう，そのリンゴ取って」といったとすると，ロボットは自分が見つけた「赤い色の丸いもの」が「リンゴ」というのだと分かる．そして，以後の会話でその物体に言及するとき「リンゴ」という言葉を使える．更に，次回，同様の物体を見つけたときには，最初から「そのリンゴを取りますか」といえるようになる．これは，まえがきで参照したRoyの研究[16]で目指していることをロボットを使っているうちに自然に行えるということになる．これについても，今後，検討していきたい．

先に，以前のシステムでは視覚情報処理の失敗をインタラクションで補うと述べた．このように視覚情報処理には失敗が多く，これを解決することが重要であることは分かっている．本論文では，簡略化発話の理解のために行動にかかわる物体を視覚情報処理で求めた．今回は，視覚情報による簡略化発話の理解の可能性を確かめることを主眼にしたので，視覚情報処理が成功するように簡単な状況に限っている．視覚情報処理が成功すれば，言語解析の部分は簡単なものなので，実験結果に示したようにシステムは想定したように動作する．実験はこのように視覚情報処理がうまくいくように環境条件を設定した．視覚情報処理に失敗すると，結局，システムは今回のシステムのもとになった以前のシステム[18],[19]と同じことになる．すなわち，例えば，「あれ取って」とユーザがいったときに，行動に関する物体が一つも見つけられなければ，図21の例のように，対象物体の特徴をユーザに聞くことになる．このようにシステム全体としては動作するが，簡略化発話理解の利点はなくなることになる．また，誤った物体を見つけてしまった場合も同様である．この場合は確認の時点で誤りに気が付き，再度，以前のシステムのレベルで依頼を理解することになり，見つけられない場合より，更に手間がかかることになる．

以上のように，今回の実験で提案の基本部分の有効性は確認できたが，実際に有用なシステムにするには，視覚情報処理の能力向上が不可欠である．以前のシステムの論文[18],[19]の中で提案した，使っているうちに場所ごとにうまくいく視覚情報処理法を獲得していく方法が有効であると考えているが，更に検討を進める必要がある．

9. むすび

人間同士の会話では，両者が視覚で情報を共有していると考えられるものについては，簡略化して発話されることが多い．ロボットに対してもこのような簡略化した表現が使えることが望ましい．そこで，人間とロボットが見たり指差したりなどの行為でかかわっている物体を検出して，簡略化された発話の理解の補助とする方法を提案した．現在は，基本的な実験システムを開発し，提案の有望性を確認したところだが，今

後，実験を進め有効性を実証するとともに，更に発話と行為の間の時間的関係など，より詳細な検討を進めていきたい．

謝辞 本研究の一部は科学研究費補助金（15017211，14350127）による．

文　献

[1] B. Graf and M. Hagele, "Dependable interaction with an intelligent home care robot," Proc. IEEE International Conference on Robotics and Automation (ICRA 2001), pp.21–26, 2001.

[2] L. Seabra Lopes and A. Teixeira, "Human-robot interaction through spoken language dialog," Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2000), pp.528–534, 2000.

[3] 渡邊　大，桝井文人，河合敦夫，椎野　努，"会話における省略とその補完，" 情処学音声言語情報処理，no.034，pp.149–154，2000.

[4] M. Schiehlen, "Ellipsis resolution with underspecified scope," Proc. 40th Annual Meeting of the Association for Computational Linguistic (ACL 2002), pp.72–79, 2002.

[5] 堂下修司，新美康永，白井克彦，田中穂積，溝口理一郎，音声による人間と機械の対話，オーム社，1998.

[6] 松尾太加志，コミュニケーションの心理学，ナカニシヤ出版，1999.

[7] Z.M. Hanafiah，中村明生，久野義徳，"視覚による音声対話の省略の補完，" 第9回画像センシングシンポジウム講演論文集，pp.433–438，2003.

[8] Z.M. Hanafiah, C. Yamazaki, A. Nakamura, and Y. Kuno, "Human-robot speech interface understanding inexplicit utterances using vision," CHI2004 Extended Abstracts, pp.1321–1324, 2004.

[9] E. Campana, J. Baldridge, J. Dowding, B.A. Hockey, R.W. Remington, and L.S. Stone, "Using eye movements to determine referents in a spoken dialogue system," Workshop on Perceptive User Interfaces, ACM Digital Library, 2001.

[10] R.A. Bolt, ""Put-that-there": Voice and gesture at the graphics interface," Comput. Graph., vol.14, no.3, pp.262–270, 1980.

[11] Y. Kuno, T. Ishiyama, S. Nakanishi, and Y. Shirai, "Combining observations of intentional and unintentional behaviors for human-computer interaction," Proc. CHI 99, pp.238–245, 1999.

[12] 松井俊浩，"おせっかいロボットとも呼ばれる事情通ロボットの計画，" bit, vol.29, no.12, pp.4–11, 1997.

[13] 松坂要佐，小林哲則，"ROBITA：グループ会話ロボット，" 人工知能学会研究会資料，SIG-Challenge-0113, pp.1–8, 2001.

[14] NEC Personal Robot Center, "Personal Robot PaPeRo," http://www.incx.nec.co.jp/robot/

[15] P. McGuire, J. Fritsch, J. Steil, F. Röthling, G.A. Fink, S. Wachsmuth, G. Sagerer, and H. Ritter, "Multi-modal human-machine communication for instructing robot grasping tasks," Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2002), pp.1082–1088, 2002.

[16] D. Roy, "Grounded spoken language acquisition: Experiments in word learning," IEEE Trans. Multimed., vol.5, no.2, pp.197–209, 2003.

[17] J.J. Gibson, The Ecological Approach to Visual Perception, Houghton Mifflin, 1979.

[18] M. Yoshizaki, A. Nakamura, and Y. Kuno, "Vision-speech system adapting to the user and environment for service robots," Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2003), pp.1290–1295, 2003.

[19] 吉崎充敏，中村明生，久野義徳，"ユーザと環境に適応する指示物体認識のための視覚音声システム，" 日本ロボット学会誌，vol.22, no.7, pp.901–910, 2004.

[20] 松本裕治，北内　啓，山下達雄，平野善隆，松田　寛，浅原正幸，"日本語形態素解析システム「茶筌」Version2.0 (manual)，" 奈良先端科学技術大学院大学松本研究室，2001.

[21] MAlib development team, http://www.malib.net/

[22] D. Coombs and C. Brown, "Real-time binocular smooth pursuit," Int. J. Comput. Vis., vol.11, no.2, pp.147–164, 1993.

[23] M. Kaur, M. Tremaine, N. Huang, J. Wilder, and Zoran, "Where is "it"? Event synchronization in gaze-speech input systemss," Proc. International Conference on Multimodal Interfaces (ICMI 2003), pp.151–158, 2003.

[24] T. Takahashi, S. Nakanishi, Y. Kuno, and Y. Shirai, "Human-robot interface by verbal and nonverbal communication," Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 1998), pp.924–929, 1998.

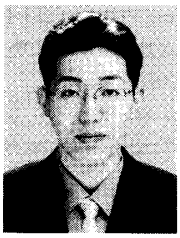
（平成16年3月25日受付，7月28日再受付）

ザリヤナ　モハマド　ハナフィア

2002埼玉大・工・情報システム卒．2004同大大学院理工学研究科情報システム工学専攻修士課程了．同年，ヒロセ電機マレーシア（株）入社．在学中，音声対話インタフェースのロボットへの応用に関する研究に従事．

山崎　千寿

2003埼玉大・工・情報システム卒．同大大学院理工学研究科情報システム工学専攻修士課程在学中．視線（共同注視）の研究に従事．



中村　明生　（正員）

1996 東大・工・精密機械卒．2001 同大大学院工学系研究科精密機械工学専攻博士課程了，博士（工学）．同年埼玉大学工学部情報システム工学科助手．複数ロボット操作システム，マンマシンインタフェース，コンピュータビジョンの研究に従事．日本ロボット学会，IEEE 各会員．

久野　義徳　（正員）

1977 東大・工・電気卒．1982 同大大学院工学系研究科博士課程了．同年，（株）東芝入社．1987～1988 カーネギーメロン大学計算機科学科客員研究員．1993～2000 大阪大学工学部電子制御機械工学科助教授．2000 より埼玉大学工学部情報システム工学科教授．工博．コンピュータビジョン，知能ロボット，ヒューマンインタフェースの研究に従事．情報処理学会，日本機械学会，日本ロボット学会，人工知能学会，計測自動制御学会，電気学会，IEEE，ACM 各会員．