ロボットへのビジョン応用の展開
ー コミュニケーションに関するビジョンー

埼玉大学大学院理工学研究科
久野義徳

ロボットが行動を行うために必要な外部情報の獲得手段としてビジョンはもっとも重要なものである。ロボットの行動としては移動とマニピュレーションがまず考えられる。したがって、そのための情報を得るためのビジョンが活発に研究されてきた。しかし、人間と共生するロボットを実現することを考えると、人間とのコミュニケーションという行動も必要になってくる。そのコミュニケーション行動にもビジョンは重要な役割を果たす。本稿では、このコミュニケーションに関するビジョンについて著者らの研究を述べる。

１．はじめに

ロボットが行動を行うために必要な外部情報の獲得手段としてビジョンはもっとも重要なものである。ロボットの行動としては移動とマニピュレーションがまず考えられる。したがって、そのための情報を得るためのビジョンが活発に研究されてきた。マニピュレーションとしては、1960 年代のハンドアイシステムの研究という、ロボット研究の当初から、ビジョンで情報を得てロボットアームを動かすことが検討されていた。また、多くの産業用ロボット向けのビジョンも広義で考えればマニピュレーション用のビジョンといえる（ロボットに限らないが、このあたりの画像処理の歴史に関しては文献[1]に詳しい）。移動についても、ビジョンを用いた多くの屋内移動ロボットが研究されてきた。また、屋外でも CMU の NavLab[2]をはじめとする多くの研究があり、これが最近の ITS につながっている。このような移動やマニピュレーションはロボットの自律的行動であり、ビジョンはそのための情報獲得手段であった。

最近では、ロボットが工場などの定められた場所だけでなく、家庭やオフィスへ入ってきて、人間と共生するようになると考えられている。人間と共生するようになると、移動やマニピュレーションだけでなく、人間とのコミュニケーションという行動が重要になる。コミュニケーションはことばで行うものと思われるが、特に対面のコミュニケーションではビジョンが重要な役割を果たしている。そこで、人間とロボットのコミュニケーションにかかわるビジョンの研究を進めている。ここでは、これに関する著者らの研究をまとめて述べる。

２．コミュニケーションとビジョン

人間同士の会話では視覚で相手の行動や周囲状況を認識していることにより、簡単な発話で相互の意図が理解できる。例えば、テーブルの上の本を取ってもらいたいとき、顔をその本の方に向けて「あれ取って」といきなり言っても、相手に通じることが多い。これは、顔を本の方に向けて、本を見ているという行動が、相手に視覚で認識され、「あれ」がテーブルの上の本を指すと相手に理解されるからである。そこで、ロボットにも人間とのコミュニケーションの際に、この能力を持たせることの実現を目指して研究を行っている。また、人間とロボットのコミュニケーションでは人間もロボットの行動を見て、そこから情報を得ると考えられる。したがって、人間の視覚に適切な情報を与える行動をロボットが行うことについても検討を行っている。ただし、どちらについても人間のあらゆるコミュニケーションの場面に対して対応できるようにはじめから考えるのは困難である。そこで、ロボットが人を見ることについては、頼んだものを取ってきてくれるなどの軽作業をしてくれる介護ロボットへの利用を想定して、人間の自然な簡略化された依頼を理解することに問題を限定している。また、人間に対して適切な行動を見せることについては、博物館や美術館でのガイドロボットの頭部の動きについて検討している。人にものを説明するときに、ずっと聞き手の方を見ていたり、ずっとものの方を見ていたりはしない。説明者はものの方を見て説明する場合でも、ときどき聞き手の方を振り返って、聞き手の様子を確認しながら、そして聞き手を引き付けるように説明を行う。このような効果的なロボットの頭部動作について検討を進めている。

今回は、以上の２つの研究について概略を述べる。さらに、コミュニケーションを通じて物体を認識する対話物体認識についても述べる。ロボットのビジョンの能力の向上は重要な研究課題だが、どのような場合でも成功するようなものの実現は難しい。人間と共生するロボットの場合は、人間とのコミュニケーション

によって、ビジョンの失敗を回復することができると考えられる。このコミュニケーションの負担が大きくてはロボットは実用にならないが、まずは、多くのコミュニケーションが必要でも、とにかく対象物を認識できるようにし、次第に負担が減るようにしていくというアプローチで研究を進めている。

３．簡略化発話の理解

先に述べたように，人間同士の会話においては、それまでに会話で触れられていない物体に対しても「あれ取って」で意味が通じることがある。このような簡略化発話の理解を一般的に扱うのは困難なので，サービスロボットへの依頼に限って検討した[3][4][5]。ロボットへの依頼は「何を」（目的語，対象物体）「どうして欲しい」（動詞）という２つの要素からなると考えた。そして，それぞれについて，

① 明確に言われている，
② 指示語（目的語の場合），代動詞（動詞の場合）が使われている，
③ 省略されている，

のどれかを判断するようにした。そして，対象物体に関して②か③の場合，人間が明確に言わなくてもわかると判断して発話したのは，それが会話の当事者（人間とロボット）の行動に関連しているからだと仮定した。実際には，その物体を指差している，手で扱っている，その物体の近くにいる，そして，その物体を見ている（視線が向いている，実装の際は概略値として顔の向き）という４つの行動を考え，それに関わっている物体を視覚で検出し，検出された物体を対象物体であるとした。動詞が②、③の場合は，対象物によって，それに対して人間がして欲しい行為は決まっているとして，解釈するようにした。

４つの行動のうち，視線以外はそれほど速く動かないので，発話された時点の行動から関連物体を検出すればよいと考えられる。しかし，視線は速く動くので，どの時点の視線上の物体を考えればよいのか検討する必要がある。そこで，以下のような実験を行った[5]。10 個程度の物体を周囲に置いた環境で，被験者にあるものを取ってもらいたいと思ってもらった上で，「あれ取って」とロボットに対して言ってもらった。ロボットは後述のものだが，画像入力以外の動作は行わない。被験者 5 人に各 10 回の試行を行ってもらい，顔の向きを調べた。

図 1 に結果を示す。10 フレームごとに，その間に顔がもっとも向いた方向を求め，全試行の平均を表示している。総和が 100%を超える場合があるが，これはロボットと物体が同じ方向にある場合などは，双方を見たとカウントしたためである。この結果から，発話の始まる前から物体を見始め，「あれ取って」という発話の最中にはほとんどの場合，対象物体を見ていることがわかる。これは，まだ予備的な実験ではあるが，この結果に基づき，発話の少し前から発話終了までの間の顔の向きを求め，その間にロボット以外の方向でもっとも顔を向けた方向の物体を探すことにした。

図 2 に、これまでの検討の結果を実装したロボットの外観を示す。このロボットは 2 組のステレオカメラを持つ。下段のステレオカメラは常に人間の方を向き、その視線や指差しの方向を求める、上段のステレオカメラは下段のステレオカメラで得られた視線あるいは指差しの方向上の物体を検出するのに用いる。視線あるいは指差しの方向の半直線上を 2 つのカメラの光軸が交わるようにカメラを動かしていき、zero-disparity filter[6]により、物体を検出する。

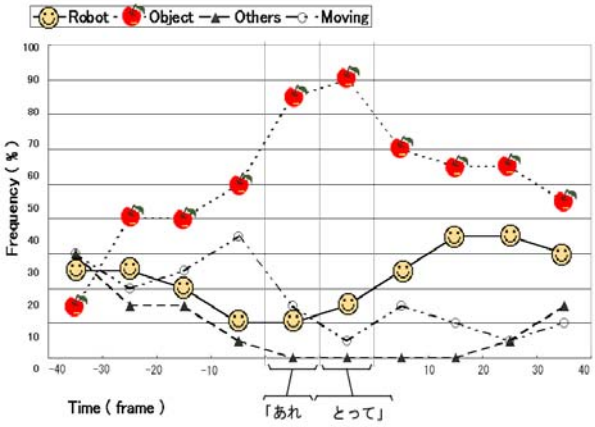


図 1　発話のタイミングと顔の向き

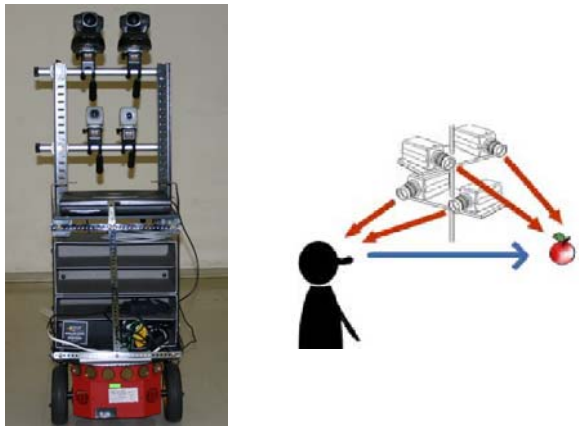


図２　簡略化発話理解ロボット

このロボットを用いて実験を行い、想定した簡略化発話の理解が視覚を用いて行えることを確認した。ただし、このシステムでは４つの行動に関連する物体を検出すればよいと仮定している。これは妥当ではあると思われるが、発話理解には他にも多くの情報を利用している可能性がある。そこで、現在、実験室に普通の家庭内のような環境を作り、車椅子使用者と介護者を想定した被験者に過ごしてもらい、そこに現れる依頼と、それがどのように理解されるかを調べている。また、老人介護福祉施設にビデオカメラを設置して、被介護者の意図を介護者がどのように理解しているかも調べている。この調査は社会学のエスノメソドロジー[7]の専門家と共同して行っている。エスノメソドロジーは人間の行動の方法について調べる研究分野だが、そこで用いられる会話分析の手法で調査を行っている。会話分析は、ビデオデータから発話前後の行動と発話の関連を分析するものである。調査はまだ途中であるが、人間同士では発話にかなりの簡略化があるのが確認されている。特に、介護施設の場合はその傾向が大きい。仮定した４つの行動に関する簡略化が実際に出現することも確かめられている。さらに、簡略化発話が理解されたかどうかを行動で互いに確認しながらコミュニケーションが進むことがわかってきた。現時点での調査結果については秋谷ら[8]に報告したが、さらに調査を進めていく。これにより、簡略化発話と関連行動の関係を明らかにし、その行動認識に基づき自然な依頼を理解できるロボットを実現していく計画である。

４．ガイドロボットの頭部動作

4．1　人間の説明場面の分析

人間が他人に展示物を説明する様子を２つの場合について調べた。一つは埼玉大学内にある古代朝鮮半島の瓦についての展示の説明である。展示内容の研究者が説明者（ガイド）になって、訪問者（すべて埼玉大学の学生）に説明する様子をビデオカメラで撮影した。１５分程度の説明を４回（別の４人)、３０分程度の説明を２人ペアーの相手に２回（２回は別人）行った。もう一つは、共同研究をしている、はこだて未来大学の山崎晶子講師のところで実施したもので、タイの写真の展示を、写真を撮影した研究者がはこだて未来大学の学生に説明しているところを記録した。３０分程度の説明を一人に対して行うものを３回（別の３人)、２人ペアーに対して行うものを２回（２回は別人）実施した。

撮影されたビデオから、ガイドの頭部の動きを社会学で使われている会話分析の方法を用いて調べた。特に、頭部を展示の方から訪問者の方に向ける部分を中心に検討した。詳細は[9]に報告したので、ここでは概略だけ述べる。分析のまとめとして、ガイドが訪問者の方へ頭部を向けたときの発話等の内容を計数した結果を表１に示す。話の切れ目になる場所は、Sacks たち[10]が発話交代部分と呼んだところと考えられる。この部分で、ガイドは顔を訪問者に向けているが、これは相手が自分の話についてきているか、何か質問がないかを確認するためだと考えられる。その他の場合では、重要な語を言うときに振り向くのは、重要な語であることを認識してもらいたい、それがわかっているか確認したいためだと考えられる。難しい語や数字を言うときも、それが伝わっているか確認したいためだと考えられる。「これ」などの指示語を使うときにも訪問者の方を向くことが多くあったが、これは、指示した方を訪問者が見ているか確認するためだと考えられる。また、このような指示語と同時に行われることが多いが、指差しなどの手のジェスチャをしたときに、振り向くことが多くあった。さらに、これは当然と思われるが、訪問者が質問をしたときには、訪問者の方を向いて質問を聞いた。

表１　ガイドが訪問者の方に頭部を向けた場合とその回数：　全部で 136 回、複数の場合があてはまる場合は、重複して計数

	回数
話の切れ目	61
重要な語を言うとき	14
難しい語や数字を言うとき	6
「これ」などの指示語を使うとき	26
手のジェスチャといっしょに	41
訪問者が質問したとき	12

4.2 簡易ロボットによる科学技術館での実験

人間の場合の分析結果に基づいて頭部を動かす簡単なガイドロボットを開発して、頭部を動かすことの効果を調べる実験を行った。このロボットは展示品の近くの人間がロボットの方を見ると、その人の顔をビデオカメラの画像から見つけて、その人に近づく（文献[11][12][13]で報告したアイコンタクトを利用)。そして、展示品の説明をするようになっている。このロボットを使って、科学技術館（東京北の丸公園）で実験を行った。図３に実験の様子を示す。磁性流体という磁石の性質をもった液体を用いた芸術作品の展示会の際に、

その作品の一つをロボットに説明させた。ロボットは、人間についての分析結果に基づき、話の切れ目や重要な語や指示語を言うときに、訪問者の方に頭部を向ける（首を振る）ようにした。この頭部動作の効果を調べるために、16 人の被験者により実験を行った。16 人を８人づつの２つのグループに分けた。一つをＡグループ、もう一つをＢグループと呼ぶことにする。Ａグループの人には、はじめに頭部を訪問者の方にずっと向けたままのロボットにより展示品の説明を受けてもらった。そして、次に頭部を動かしたロボットの説明を受けてもらった。Ｂグループの人には、この逆の順序で説明を受けてもらった。なお、２回の実験でのロボットの動作の違いなどは被験者には告げなかった。

図４は実験結果をまとめたものである。この図は、実験に参加した個人ごとに、首を振らずに固定したロボットに対して、説明の間にロボットの方に頭部を向けた数を横軸に、説明の適当なところで首を振るロボットに対して、同様にロボットの方を向いた数を縦軸に示したグラフである。○はＡグループの人、△はＢグループの人に対する結果であることを示す。頭部を動かすロボットに対しては、人間の方も頭部を動かす回数が有意に増えていることがわかる($p<0.01$)。

図３　科学技術館での実験

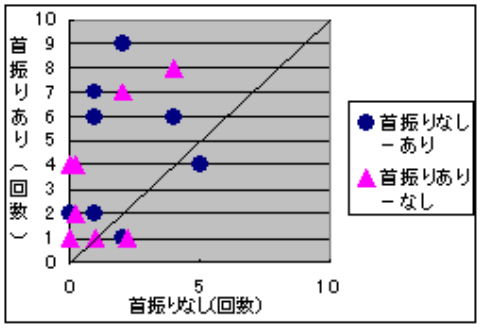


図４　ロボットの首振りがある場合とない場合に対する人間の頭部動作の回数

回数が増えれば良いとは必ずしもいえないが、回数が多いほうがロボットとのかかわりが強かったということで、ロボットの説明がより訪問者をとらえていたと考えられる。ただし、この実験では頭部を動かさないロボットと比較したので、動かし方が良かったかどうかは断定できない。そこで、現在、ヒューマノイドロボットを用いて、頭部を提案のように動かす場合と、ランダムに動かす場合の比較実験を行っている。

５．対話物体認識

これまでに述べた研究はコンピュータビジョン技術の利用法に関する研究であり、その中で用いられるコンピュータビジョン技術に関しては、既存技術を利用している。また、３のシステムは簡単な環境で対象物体も顕著なものに対して動作を確認しただけである。実際に有用なサービスロボットを実現するためには、これまで述べた応用システムに加えて、家庭内のような一般的環境で、依頼対象になる物体を認識できるコンピュータビジョンの技術が必要である。2 章で述べたように、これを人間との対話を通じて、人間からの支援を得て実現することについて研究を進めている。

コンピュータと対話しながら物体を特定したりシーンを理解したりすることについては Winograd の古典的研究[14]以来、多くの研究がある。しかし、これらの研究では、シーン内の物体は形や色などで記述されており、記号として扱えるようになっている場合が多い。実際にロボットを扱う研究では、実世界が信号レベルから考慮されるようになってきている。例えば Inamura らの研究[15]では獲得情報のあいまい性を考えてマルチモーダルな情報から人間の意図を理解している。しかし、視覚の部分に限れば対象物体は簡単なものに限定されている。それに対し、著者らのグループでは物体認識というコンピュータビジョンの問題を中心に検討を進めている。

5.1 物体認識システムの構成

物体認識システムの詳細に入る前に、著者らの物体認識の問題に対する考え方を述べる。物体認識の問題は多くの課題に分けられるが、ここでは以下のように分類して考える。

A. 対象に対する先験的知識がある場合

① 概念レベルの認識：物体に与えられた一般的な名前で認識できる。例えば、どんな「本」でも「本」と認識できる。

② 特定物体レベルの認識：特定の物体を事前に見たことがあり、その特定の物体を認識できる。例えば、特定の雑誌の特定の号を以前に見ており、今回、その雑誌を認識できる。

実際には、対象の一般性に関して①と②の間にさらに多くのレベルの設定が考えられる。

B. 対象に関する先験的知識がない場合

これは再認(recognition)という意味では認識ではないが、ある一つの物体を他の物体と切り分けて認知できる能力である。

人間の場合はほとんどの場合、A①のレベルで認識できる。知らない物体については、その名前を言われても認識はできないが、B のレベルの認識は通常は問題なくできる。ここで主に考えている「ものを取ってくる」というような依頼において、人間の場合でも頼まれた方が対象物がわからず聞き返すことはある。考えられる場合としては、３で扱ったような簡略化発話では対象がわからなかった場合がまずあげられる。「あれ取って」と言われても「あれ」が何かわからなかったような場合である。この場合は、人間の場合は「そこの本」というように、物体名を告げるなどの詳細情報を与えると考えられる。それにより A①の能力により物体を認識できる。ときには、物体名を与えられてもわからない場合もある。これには、まれではあるがその物体名を知らなかったり、その物体名の中でのバリエーションを知らなかった場合、その物体名の物体が複数あり、そのどれかわからない場合、視野にあるのにたまたま目に入らず見つけられない場合がある。この場合は物体の属性や位置に関する情報を依頼者がさらに与えることで、普通は認識できる。物体名を知らない場合でも、B の場合に対する能力で切り分けて認知した物体の中で、依頼者の属性や位置情報に合うものを認識することができる。

以上のような人間の場合に対し、ロボット（コンピュータビジョン）では、A①のレベルの認識能力は現状ではきわめて限定的なものである。A②のレベルについてはかなり進歩してきているが、様々な条件変化に対して必ず動作が保障できるレベルではない。したがって、A①、②に対応しようとした視覚をもったロボットでは認識の失敗を避けられない。そこで、そのような場合を救うために、対話物体認識を検討している。A①、②に対応した物体認識が失敗したということは、それに用いた先験的知識が有効でなかったということになる。対話により、知識を補えば認識できる場合もあるが、一般的にこの場合を救おうとするなら、B の先験的知識がない場合に対応する物体認識を準備しておく必要がある。しかし、ここでも人間とコンピュータの能力の大きな違いが問題になる。B の場合へ対応できるということは、コンピュータビジョンでいえばセグメンテーションができるということである。人間にとっては、ふつう、これは問題ない。したがって、先に述べたように物体名を知らない場合でも、属性や位置の情報を与えてもらえば対象を認識できる。しかし、コンピュータビジョンでは完全なセグメンテーションは難しい。そこで、ここにも対話による人間からの支援を考える。

研究の流れとしては、きわめて限定された場合であるが A①のレベルの視覚の失敗を記号レベルの世界の中で補うシステムから対話物体認識にとりかかった。これを 5.2 で述べる。それから、5.3 で述べるセグメンテーションが完全な場合への対応に進んだ。そして、5.4 で述べるセグメンテーションが不完全な場合に、対話でそれを補うシステムへと研究を進めている。

ただし、以下で述べる対話物体認識だけでサービスロボットの視覚を構成しようと考えているのではないことに注意して欲しい。実際のロボットとしては、A①、②に対応するシステムを持ち、それらが失敗したときに対話物体認識を使用するという階層的なシステムを考えている。特に、先験的知識がなく、セグメンテーションもうまくいかないという場合に対応する 5.4 の研究を進展させたものを対話システムの中でも最下層におき、手間はかかるかもしれないが、対象物がわからないことはないというシステムを目指している。

5.2 記号レベルのシステム

対話を通じた物体認識として、最初はまず記号レベルでの支援から検討を開始した[16]。人間同士では対話を通じてお互いの共通理解を目指していると考えられる。したがって、対話においては、何がわかっていて、何がわかっていないかを相手に伝えることがコミュニケーションを進める鍵になる。それが伝われば、相手はわかっていないことに関する情報を与えてくれることが期待できる。この考えに基づいて、サービスロボットの視覚システムを開発した。

このシステムでは、物体は色や形の属性で記述する。システムでは「○○を取って」という依頼が来ると、知識ベースの中からその物体の属性を調べてそれをゴールに設定する。一方、属性検出の画像処理を起動する。実装された属性は色、形、個数の３つだけである。検出された領域の属性とゴールを照合する。もし、すべての条件を満たす領域があれば、それが対象かどうか人間に確認する。一部しか照合するものがない場合は、照合する部分は肯定で、照合しない部分は否定にして、認識結果を言葉で説明する。それに対して人間が属性に対しての発話を行えば、ゴールの属性をそれに変更して判断処理を繰り返す。

このシステムでは以下のような対話物体認識が可能である。例えばリンゴは「色：赤；形：円」というように記述されているとする。ロボットは「リンゴを取って」という依頼が来たとき、画像中から赤い丸い領

域が検出できたら、人間に、それが目的物かどうか確認する。検出できない場合は、画像処理の結果を人間に伝える。例えば、シーンには黄色いリンゴしかなく、人間が頼んだのはその黄色いリンゴだったとする。この場合、ロボットは黄色で丸い領域は検出できたが、赤い丸い領域は検出できない。そこで、「赤色でない丸いものを見つけました」と音声で人間に言う。それを聞くと人間はロボットが黄色いリンゴの存在を知らないということがわかり、「黄色だよ」というような色に関する情報をロボットに伝えてくれることが期待できる。このような発話があると、ロボットはゴールの色の部分を黄色にして、リンゴを検出できる。

実際のこの研究は従来の記号レベルの研究と同様なもので、コンピュータビジョンは完全に物体についての記号レベルの記述を与えてくれるものと仮定している。5.1 で述べたが、A①のレベルの視覚が備わっていると仮定していることになる。しかし、実際には A①といっても、利用している先験的知識は赤い丸いものがリンゴ、四角いものが本というレベルのものであり、シーンの中にも物体が２～３個しかないと仮定している。したがって、これだけでは限定した場合にしか動作しない、トイシステムにすぎない。しかし、5.1 で述べたように、階層的なシステム構成を考え、失敗したら下位にいけばよいとできれば、限定した場面でしか動作しなくても意味があることになる。ただし、現時点のシステムでは限定が強すぎるので、A①、②レベルのさらに能力の高い物体認識と組合せることを検討する必要がある。以上のように、このシステム自体ではまだ技術的課題は多いが、「わかっていることとわからないことを伝えると人間からわからないことについての情報が得られる」という考え方は、以降の研究につながっている。

5.3　物体の特定

5.2 の研究では対象シーンが簡単ということで暗黙のうちに、画像中には対象物体の他には少数の物体しかないと仮定していた。しかし、実際のシーンでは画像をセグメンテーションして、その中から対象物体を検出しようという場合、セグメンテーション結果の中に多数の領域が含まれるのが普通である。そこで、多数の領域の中からどれが対象物体かを対話により特定する方法を検討した[17]。これも領域の属性で処理を考えており、まだ記号レベルでの扱いともいえるが、対話を生成する際には画像処理のことを考慮に入れている。

ここで扱う問題は、画像中の多数の物体の中から人間が決めた物体を人間に対象について質問をすることにより情報を得て特定することである。質問に対する答えから該当する候補を絞っていき、物体を特定するわけだが、結果を得るまでの質問の数ができるだけ少なく、また、それぞれの質問が人間にとって答えやすいものであることが望まれる。研究課題は、画像が与えられたとき、このような質問を生成する方法を検討することである。

まず、画像に対して、色情報に基づくセグメンテーションを行う。ここでは、セグメンテーション結果の各領域が一つの物体に対応していると仮定する。したがって、どの領域が対象物に対応するかを決定することが課題になる。各領域について、色、形などの特徴を求め、その特徴に関して人間に質問を行う。質問の生成に際しては、特徴の性質を考慮して、人間に答えやすく効率的に対象を絞り込める質問を作成する。ここでは、以下の４つの性質を考える。

① 語彙の豊富さ
② 分布により影響を受けるか
③ 唯一性
④ 絶対的か相対的か

①はその特徴を表す語彙がたくさんあり、人間が言葉でその特徴を表すことが簡単にできるかどうかということである。色はこれに当てはまる。形も語彙は豊富だが、言葉で表しにくい形も多いので、これに当てはまらないと考える。②は周囲に他の物体が存在するときに影響を受けるかどうかということである。例えば位置に関する表現は、物体の分布により、使えるものが変わってくる。それに対し、色ではそういうことはない。③は同じ値を持つものが他にあるかということである。④は大きさの大小などは他のものの存在により変わる可能性があるので相対的なのに対し、形はそういうことがなく絶対的であるというようなことである。 今回のシステムでは、特徴としては、色、形、大きさ、位置の４つしか扱っていないが、それらの特徴についてここで述べた性質をまとめたものを表 2 に示す。

セグメンテーション結果の領域の特徴の分布と特徴の性質から、答えやすく効率的な質問を生成する。質問の形としては(1)「何？」という質問（例えば「どんな色ですか」というような質問）、(2)「はい・いいえで答えられる質問」、(3)「A、B、Cのどれですか」というような質問が考えられる。対象物体の候補を少数に絞り込む効率という観点からいえば、(1)の質問がよいことになる。しかし、(1)の質問は人間には答えにくい場合がある。(2)の質問は(1)と反対の性質をもつ。したがって、どういう特徴に関して、どちらの形の質問をするのが良いかを考えることになる。なお、今回のシステムでは(3)は特別の場合にしか用いていない。

表２　特徴の性質

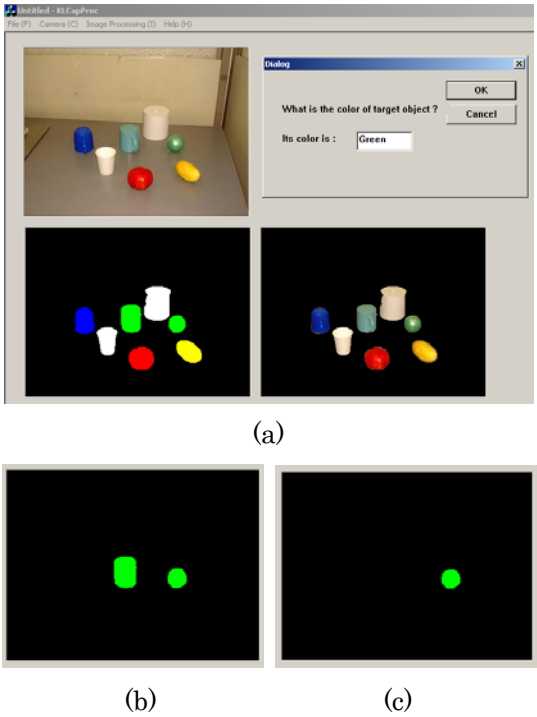
Characteristic	Color	Size	Position	Shape
Vocabulary	rich	-	rich	-
Distribution	-	-	dependent	-
Uniqueness	-	-	unique	-
Absolute/Relative	absolute	relative	relative	absolute

紙数の関係で質問文生成の方法の詳細は省略するが、概略の方針は以下のようなものである。まず、それぞれの特徴について、分布を調べる。例えば色特徴は７つのクラスに分けているが、それぞれに何個の領域があるかを求める。そして、分布が一番分散している特徴を求める。その特徴が語彙が豊富なものであれば、その特徴に関して「何？」という質問を行う。例えば、領域の色が多くの色のクラスに分かれていれば、色は語彙が豊富な特徴なので、「何色？」と聞く。こう聞かれても、語彙が豊富な特徴に対してなら人間は容易に答えられる。ある特徴についてある特定の値をもつ領域が多数ある場合には、それについては「はい・いいえで答えられる」質問を行う。例えば、四角形の領域が多く、少数が円の場合は、「四角形ですか」と聞く。そして、対象の数が絞られてきたら、位置関係（右・左等）の質問を用いることも検討する。 図５に対話の例を示す。この例では、最初にロボットは「物体は何色か」と聞く。人間が「緑」と答えると図５(b)の物体が候補として残る。そこで、ロボットは「左の物体か」と聞く。人間が「違う」と答えたので、ロボットは図５(c)に示す物体を対象物と認識する。

４つのシーンで 10 人の被験者に実験者が意図した物体を質問をして当ててもらう実験を行い、提案システムの場合と比較した。その結果、提案システムが質問数で多くなる例はなかった。したがって、生成した質問は効率的であると判断できる。また、提案システムでは語彙が豊富な特徴でなければ「何？」という質問は行わないので、答えやすい質問が生成されていると考えられる。

5.4　対話によるセグメンテーション

前節の方法により、多数の物体の中から目的の物体を対話により特定できるようになった。しかし、この方法では、画像のセグメンテーションに誤りがなく、一つの領域が一つの物体に対応していると仮定している。画像特徴の性質を考慮して質問文を生成するという基本的な方法は、一般的に使えるものとして重要な成果と考えられるが、実際のシーンに対しては、この仮定が成り立つことは期待できない場合も多い。物体同士に重なりがあったり、一つの物体が複数の色で構成されていたりすれば、この仮定が成り立たなくなる。さらに、そのような場合でなくても、セグメンテーションが完全でない場合もある。そこで、対話を通じてセグメンテーションを修正する方法を検討した[18][19]。



(a)

(b)　　(c)

図５　対話物体認識の例　(a) 原画像（左上)、カラーセグメンテーション結果（左下)、切り出された物体(右下)　(b) 最初の質問の後に残った物体　(c) 最後の質問の結果

最初に提案した方法[18]の基本的な部分は以下のようなものである。領域の境界となる可能性のある部分の両側の詳細な特徴を調べる。その結果、境界の両側の部分が同一物体か、異なる物体か確かに判断できる場合以外は、人間にどちらであるか尋ねる。

しかし、この方法には２つの問題点がある。一つは、境界候補ごとに人間に判断をあおぐので対話数が多くなってしまうことである。もう一つは、コンピュータの方で確かと判断した場合でも間違いがあり得ることである。これを防ぐにはすべての境界候補に対して人間に質問をすればよいが、これではさらに対話数が増えてしまう。

そこで、以下のように改良した方法を考案した[19]。まず、境界候補の詳細特徴を調べる。そして、特徴の値を利用して、可能性の高い順にセグメンテーション結果の候補を生成する。その際、必要ならばさらに別

種の詳細特徴を調べる。それから、もっとも可能性の高いセグメンテーション結果から人間に説明して、正しいか確認する。この説明の際には、最初に解釈した物体の数を伝え、人間に確認を求める。この数が違えば、次の候補に進む。数が同じ場合は、どういう物体があるかセグメンテーション結果の詳細説明を行う。人間がそれに同意すれば解釈が正しかったことになる。違いを指摘された場合は他のセグメンテーション結果候補のうち、その数のものの検証に進む。なお、以上のプロセスの際に、特徴量からでは判定ができない境界候補があれば、それについては、最初に提案した方法と同様に、その境界についての判断を人間に聞く。現時点の実装では、その境界部分をディスプレイ上に示し、その両側が同一物体か別の物体か聞く（これに関しては 5.5.1 参照）。

実装したシステムでは最初に調べる境界部分の詳細特徴に reflectance ratio を用いた[20]。reflectance ratio は境界の両側が同一面上にあれば境界上で同じ値をとる。そこで、境界線をはさむ 2 点間のこの値を求め、その分散を計算する。分散が小さければ、境界の両側は同一物体の可能性が高い。逆に大きければ、２つの領域は画像上では隣接しているが空間的には離れた別の物体である可能性が高い。しかし、２つの物体が密着している場合は、当然、分散は小さくなる。他にも、いろいろな場合があり、この値による判断が絶対というわけではないが、分散が実験的に定めた上限しきい値以上なら２つの物体、下限しきい値以下なら同じ物体と判断して仮説生成に進む。この２つのしきい値の間の場合は、次に示す特徴を調べる。

reflectance ratio で判断できない場合には、境界線を通る線分上の輝度プロファイルを調べる。照明条件や物体の反射特性など、いろいろの条件が関わるが、shape-from-shading の研究で示されたように、輝度の変化は３次元の形の変化に対応する場合がある。そこで、境界の両側の輝度プロファイルを調べ、それに直線あるいは２次曲線を当てはめる。そして、境界をはさんだ両側が３次元世界で接続している面上にあり得るパターンかどうか判定する。判定法を表３に示す。面が接続していても、例えば色がそこで変わるなどで輝度の値が連続しているとは限らないので、この判定法では輝度プロファイルの形という定性的な指標で連結性を判定している。この判定も絶対に正しいというわけではないが、境界の両側が同一物体かどうかの可能性についての判断の指標として仮説生成に用いる。

図６に実験結果の例を示す。図中の数字は reflectance ratio の分散だが、中央付近の黄色と赤色のボールの境界の分散が 0.0013 と小さいため、最初は同一物体だと判定し、「物体数は５つですか」と人間に聞く。人間が「６」と答えると、図６(a)に白線で示す部分の輝度プロファイルを調べる。その結果が図６(b)、(c)である。このようなプロファイルの組合せは表２の中の同じ物体と判定される中にない。したがって、この２つの部分は別の物体であり、物体数は６であるという仮定を得る。物体数が６ということはすでにわかっているので、「２つの赤い物体、２つの黄色い物体、１つの青い物体、１つの緑の物体がありますか」と仮定の内容を人間に説明して確認を求める。

表３　輝度プロファイルのパターンによる判定

Intensity profile of region 1	Intensity profile of region 2	Decision
Line:	Line:	Same object
Line+:	Line-	Same object
Line-:	Line+:	Same object
Curve:	Curve -:	Same object
Curve +:	Curve -:	Same object
Curve -:	Curve +:	Same object
Other combinations		Different objects
Too small or big region (s)		Unknown

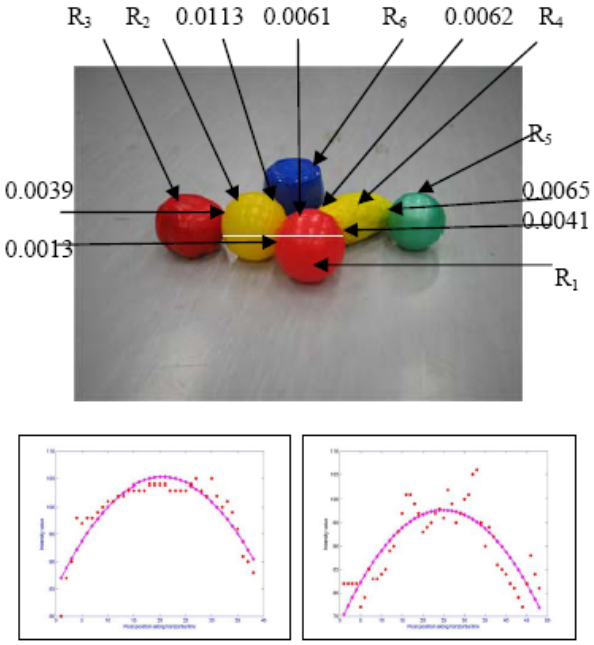


図６　対話によるセグメンテーションの例
(a)画像データ(上)、(b)黄色物体上の輝度プロファイル(左下)、 (c)赤色物体上の輝度プロファイル(右下)

この例のように仮説が違っている場合、輝度プロファイルを調べないで判定を下していた境界があれば、輝度プロファイルを調べる。物体数を多く考える必要がある場合には、reflectance ratio の分散が下限しきい値より小さくて、同一物体と判定した部分を対象に調べる。逆の場合は上限しきい値を超えた部分を調べる。複数ある場合は、しきい値に近い方から調べる。

この例を含めて、１７種類の物体から２～６個を選び、80 シーンを構成して実験を行った。実験では、全部で 335 個の境界があったが、そのうちの 81%は自動判断を正しいとして仮説を生成し処理を進め、確認以外の対話は必要なかった。残りの 19%については、対話による支援が必要であった。

この方法の基本的な考え方をまとめると以下のようになる。まず、対話システムといっても、できるだけ人間の負担を減らすためには、正しい解釈が自動的にできるのが望ましい。そこで、多くの詳細な特徴を調べるようにする。さらに、人間が簡単に確認できるように、境界候補ごとでなく、全体の解釈の仮説を立てる。一方で、間違いの可能性を考え、最終的には人間に解釈結果を示して確認してもらう。実際に、これをシーン全体に行うと実用的でないおそれがあるが、ここでは、サービスロボットの視覚として、人間の意図した物体を検出するのが目的であり、その目的物体周辺についてセグメンテーションができればよいので、適当な対話数で目的が達成できると考えられる。

5.5　対話物体認識の課題

物体の特定とセグメンテーションを対話を通じて行う研究について述べたが、これらはまだ研究の初期段階で課題が多く残されている。ここでは、そのうちの主なものについて議論する。

5.5.1　空間関係の表現

研究課題としては、まず、5.3 と 5.4 で述べた物体の特定とセグメンテーションの両者を統合する必要がある。ここで考えている視覚はシーンを理解することでなく必要な物体を検出することである。したがって、先に述べたように、セグメンテーションの修正は、特定の物体の検出に必要な部分について行えばよい。対話を通じて、目的の物体のある部分を限定していき、物体の特定に必要な部分に対してだけ、対話によるセグメンテーションを行えばよい。しかし、ここに大きな問題がある。実は、5.3、 5.4 の個々のシステムの場合にも同じ問題が存在する。それは、人間とロボットがどの部分について対話しているのかを、どのように理解しあうかという問題である。5.3、 5.4 の現状のシステムではロボットが処理対象画像をディスプレイに表示して、この問題を回避している。すなわち、ロボットが発話する際に、発話の対象になっている、例えば境界候補などを、ディスプレイの上で人間にわかるように表示している。これは実用的な方法ともいえるが、やはり、音声対話だけで問題を解決したい。

この解決の方法として reference system の利用を検討している[21]。reference system とは、空間表現の枠組みを考えて、それで物体の位置を指定しようというものである。例えば、ある物体を基準にとって、それに対する位置関係で他の物体を示すようなことである。文献[21]では物体に重なりのある場合には、形や大きさなどの特徴が画像から得られないので、それを補うものとして、reference system の利用を提案したが、話をしている部分を会話の両者で共通理解するための方法として、重要な働きをするものとして研究を進めている。

reference system については、人間がどのように空間を認知して表現するかという観点から心理言語学などで多くの研究がある。Levinson[22] は人間の reference system を intrinsic、 relative、 absolute の３つに分けている(intrinsic はものの名を言えば、前後などが決まっているものを利用する方法、relative はＡから見てＢのＣにあるというようなもの、absolute は東西など絶対的に決まっている指標を利用するもの)。文献[21]では、シーンの中から顕著な物体を選び、それを基準にして、ロボットあるいは人間から見て、その基準物体からどこにあるものという形で、relative system を利用することを提案した。また、複数の物体が固まっている場合には、固まり全体を一つの対象と考え、固まりの中のどこという形での空間表現も提案した。これは Tenbrink らが移動ロボットの行き先の指令の際に考えた group-based reference system[23] に相当する。対話の際に、人間とロボットで共通の認識ができた物体があれば、それを基にした reference system で位置関係を表現し、別の物体が共通認識できる。これを順次行えば、人間の意図した物体にたどり着けると考えられる。例えば、シーン中に赤い物体が一つしかなければ、赤い物体と言えば、簡単に共通認識ができる。こういった物体を手始めにして対話を進めていけばよい。 文献[21]では基準物体の選択法や利用などについて簡単なものを提案したが、さらに研究を進める必要がある。

5.5.2　対話と学習

ここでは対話を通じた物体認識について述べたが、最初のうち、対話を使わなければならないのはよいとしても、いつも同じように対話が必要では人間は使う気にならなくなる。しかも、会話は色や形や位置関係

などの属性に関する語彙による会話である。人間にとっては、やはり物体の名称を使った会話が自然である。提案したような対話物体認識システムを使った場合でも、人間は対象物の名称をどこかで言うと考えられる。そこで、その物体を属性に関する語彙の対話で認識できたら、その物体と人間の使った名称を対応付けて、次回からは名称で指示されても認識できるようにすることが課題として考えられる。もちろん、次回に名称による指示で認識を試みて、できない場合には対話を用いればよい。これは行動を通じながら言葉の意味を理解させようという Roy らの研究[24]と通じるものであるが、実用的なロボットの実現のために重要な研究課題である。

この学習の問題については、実際の検討はまだあまり行っていない。関連するものとしては、成功した画像処理法を場所に結び付けて記憶しておくことを提案した程度である[25]。アフォーダンスの提唱者の Gibson は視覚認知の対象の分類の中で、付着対象と遊離対象というカテゴリーをあげている[26]。付着対象は環境に固定されていて動かないもの、遊離対象はそうでないものである。ここで考えているサービスロボットに関していえば、人間が取ってきてもらいたいようなものは遊離対象である。そして、付着対象は、家具などであり、そこに遊離対象が置かれるなどしている。そこで、遊離対象の認識のときに、関連する付着対象を人間に言ってもらい、そこで成功した物体認識のための画像処理法やそのパラメータを付着対象に関連付けて記憶しておく。次に、その遊離対象が言及されたとき、関連する付着対象として記憶した付着対象が指示されれば、記憶されている情報を最初に使って認識を試みる。もちろん、それでうまくいかなければ、対話で支援を求める。これは、照明条件など変わる可能性もあるが、付着対象により環境が規定されるという考えに基づいている。また、付着対象は動かないので、一度、どこにあるかがわかれば、それ以降は視覚で認識する必要はない（ロボットが自己の位置を知っている必要はあるが）。しかも、人間は付着対象をその名称で指示できる。不確実な視覚情報処理を避けて、環境情報を得られるという点でも、付着対象の利用は有効である。5.5.1 で述べた空間関係の表現と組合せれば、さらに利用範囲が広がると考えられる。このように付着対象の活用は有効だと思われるが、学習に関しては多くの課題があり、今後、検討していきたい。

5.5.3　セグメンテーションと認識

ここでは、セグメンテーションをしてから認識という手順を考えている。より正確に言えば、セグメンテーションで個々の物体に対応する領域を得て、その領域の属性を調べて人間の意図した物体を検出している。先験的知識が全くない場合には、このようにセグメンテーションが基本的には必要になる。しかし、先験的知識が利用できる場合にはセグメンテーションをしなくても、知識に合う部分の検出をすることが認識になるという方法も考えられる。このような物体認識としては、SIFT アルゴリズム[27]をもとにした方法などが提案され、複雑な背景での動作などの有効性が示されている。ここで提案の方法とこのような物体認識法との統合も今後考えていきたい。

５．まとめ

ロボットへのビジョンの新しい応用として、コミュニケーションに関するビジョンについて述べた。対面のコミュニケーションではビジョンで相手の行動や周囲の状況の情報を得ていることがコミュニケーションを円滑に進めるために重要である。ここでは、それに関する研究の第一歩として、ビジョンを利用した簡略化発話の理解について述べた。また、人間が相手の行動を見ているということは、ロボットも人間の視覚に適切な行動を見せる必要がある。これについては、ガイドロボットの頭部動作について述べた。また、コミュニケーションを利用するという観点から、物体が認識できない場合に人間との対話を通じて物体を認識する研究について述べた。

これらの研究は、まだ個々の部分についても研究すべき課題は多いが、それらの研究を進めて、統合したシステムの実現を目指していきたい。また、これらの研究は人間に深く関わっている。そこで、工学の技術面からだけの検討だけでなく、人間の性質を調べて研究を進めようということで、社会学の専門家と共同研究を進めている。分野融合的な研究を進めることで、人間に真に有用な技術を開発していきたい。

本研究の一部は総務省戦略的情報通信研究開発推進制度、日本学術振興会人文・社会科学振興プロジェクト研究「日本の文化政策とミュージアムの未来」、科学研究費補助金(14350127、 18049010)による。

参考文献

[1] 江尻正員：画像処理の歴史, ディジタル画像処理編集委員会 監修, ディジタル画像処理, 財団法人画像情報教育振興協会, pp.332-341 (2004).

[2] The Robotics Institute, Carnegie Mellon University, http://www.ri.cmu.edu/labs/lab_28.html

[3] Hanafiah Z.M., Yamazaki C., Nakamura A., and Kuno Y.: Human-Robot Speech Interface Understanding Inexplicit Utterances Using Vision, *CHI2004 Extended Abstracts*, pp.1321-1324 (2004).

[4] ザリヤナ・モハマド・ハナフィア, 山崎千寿, 中村明生, 久野義徳：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, 電子情報通信学会論文誌, Vol.J88-D-II, No.3, pp.605-618 (2005).

[5] 山崎千寿, 久野義徳, 中村明生：人間とロボットのコミュニケーションにおける顔の向き情報の利用, 画像の認識・理解シンポジウム(MIRU2005), CD-ROM (2005).

[6] Coombs, D. and Brown, C.: Real-time Binocular Smooth Pursuit. *International Journal of Computer Vision*, Vol.11, No.2, pp.147-164 (1993).

[7] 山崎敬一編：実践エスノメソドロジー入門, 有斐閣(2004).

[8] 秋谷直矩, 丹羽仁史, 久野義徳, 山崎敬一： 福祉ロボット開発のための依頼のプロセスに関する基礎的考察, 電子情報通信学会技術研究報告 福祉情報工学, Vol.105, No.684, pp.35-40 (2006).

[9] 森山正太, 関口博之, 坪田寿夫, 山崎敬一, 久野義徳, 山崎晶子, 解説時の視線のエスノメソドロジー的分析に基づくガイドロボット： 電子情報通信学会技術研究報告 人工知能と知識処理, Vol.105, No.639, pp.29-34 (2006).

[10] Sacks, H., Schegloff, E., and Jefferson, G.: A Simplest Systematics for the Organization of Turn-taking in Conversation, *Language*, Vol.50, pp.696-735 (1974).

[11] Miyauchi, D., Sakurai, A., Nakamura, A., and Kuno, Y.: Active Eye Contact for Human-Robot Communication, *CHI2004 Extended Abstracts*, pp.1099-1102 (2004).

[12] Miyauchi, D., Nakamura, A., and Kuno, Y.: Bidirectional Eye Contact for Human-Robot Communication, *IEICE Trans. Inf. &. Syst.*, Vol.E88-D, No.11, pp.2509-2516 (2005).

[13] Kuno, Y., Nakamura, A., and Miyauchi, D.: Beckoning Robots with the Eyes, *Proc. International Workshop on Intelligent Environments*, pp.260-266 (2005).

[14] Winograd, T.: *Understanding Natural Language*, Academic Press, New York (1972).

[15] Inamura, T., Inaba, M., and Inoue, H.: Dialogue Control for Task Achievement based on Evaluation of Situational Vagueness and Stochastic Representation of Experiences, *Proc. International Conference on Intelligent Robots and Systems*, pp. 2861-2866 (2004).

[16] 高橋拓弥, 中西 和, 久野義徳, 白井良明：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, インタラクション'98 講演論文集, pp.161-168 (1998).

[17] Kurnia, R., Hossain, M.A., Nakamura, A., and Kuno, Y.: Generation of Efficient and User-friendly Queries for Helper Robots to Detect Target Objects, *Advanced Robotics*, Vol.20, No.5, pp.499-517 (2006).

[18] Hossain, M.A., Kurnia, R., Nakamura, A., and Kuno, Y.: Interactive Object Recognition System for a Helper Robot Using Photometric Invariance, *IEICE Trans. Inf. &. Syst.*, Vol.E88-D, No.11, pp.2500-2508 (2005).

[19] Hossain, M.A., Kurnia, R., Nakamura, A., and Kuno, Y.: Interactive Object Recognition through Hypothesis Generation and Confirmation, *IEICE Trans. Inf. &. Syst.*, Vol.E89-D, No.7, pp. 2197-2206 (2006).

[20] Nayar, S.K. and Bolle, R.M.: Reflectance based Object Recognition, *International Journal of Computer Vision*, Vol. 17, No. 3, pp. 219-240 (1996).

[21] Kurnia, R., Hossain, M.A., Nakamura, A., and Kuno, Y.: Using Reference Objects to Specify Position in Interactive Object Recognition, *Proc. International Conference on Instrumentation, Communication and Information Technology*, pp.709-714 (2005).

[22] Levinson, S. C.: Frames of Reference and Molyneux"s Question: Crosslinguistic Evidence, Bloom, P., Peterson, M., Nadel, L., and Garrett, M., Eds., *Language and Space*, MIT Press, Cambridge, MA, pp. 109–169 (1996).

[23] Tenbrink, T. and Moratz, R.: Group-based Spatial Reference in Linguistic Human- Robot Interaction. *Proc. The European Cognitive Science Conference* (2003).

[24] Roy, D.K. and Pentland, A.P.: Learning Words from Sights and Sounds: A Computational Model, *Cognitive Science*, Vol.26, No.1, pp.113-146 (2002).

[25] 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).

[26] Gibson, J.J.: *The Ecological Approach to Visual Perception*, Houghton Mifflin (1979). (古崎 敬 他訳：生態学的視覚論, サイエンス社 (1985)).

[27] Lowe, D.: Distinctive Image Features from Scale-Invariant Key-points, *International Journal of Computer Vision*, Vol.60, No. 2, pp. 91-110 (2004).