2005－CVIM－147 (10)
2005／1／20

コンピュータビジョンの課題

白井良明
大阪大学大学院工学研究科電子制御機械工学専攻
〒565-0871 大阪府吹田市山田丘2-1

コンピュータビジョンの課題を，筆者のこれまでの研究を参照しながら述べる．それらは，人工知能としてのビジョンの課題，入力情報の拡大，CVは何を仮定することができるか，人との協力をめざすインタラクティブビジョンなどである．これ以外の難しい課題も示している．これまで，それぞれの課題の解決を試みたが，多くの問題が残されている．

Human Tracking in Complex Environment

Yoshiaki Shirai
Dept.of Computer-Controlled Mechanical Systems, Osaka University
2-1 Yamadaoka,Suita,osaka 565-0871,Japan

This paper describes problems of computer vision referring to author's works during 30 years. They are "Birth of vision in AI", "Extension of input information", "What are assumed", "Interactive vision" and so forth. Although we have attempted to solve difficult problems, there is still much to be done.

1 はじめに

コンピュータビジョンは，パターン認識を源流として，揺籃期を人工知能の環境で育ったといえよう．前者は実用化をめざし，画像理解は人間の知能に迫ることをめざしていた．その概要は以下のようである．

1. パターン認識のみ（1960年代中期まで）
2次元の文字や図形が対象で，パターン認識理論が中心．郵便番号の読み取り装置が実用化．

2. 初期（1960年代後期）
顔写真認識が始まる．人工知能で，次元を対象とした認識が始まる．

3. 発展期（1970年代前期，中期）
対象が，機械部品，風景に拡大，知識の利用が重要視され，コンピュータビジョン，画像理解という言葉が出現．

4. 円熟期（1970年代後期から80年中期）
距離画像，動画像が扱われるようになり，3Dモデルやオプティカルフローの利用が始まる．

5. 転換期（1980年代後期から90年代中期）
対象の拡張は一段落し，基礎的な理論が盛んになる（反射モデル，Factorization法，エピポーラ幾何学など）．理論は新しいものばかりでなく，固有空間法のようなリバイバル，ニューラルネット，HMMのような既存ものがある．

6. 人間期（1990年代後期から現在）
研究分野が広がっているのでいちがいにはいえないが，一つの傾向は，人間と密接に関係するようになったことであろう．すなわち，
- 対象：顔，ジェスチャーや行動など
- 応用：ヒューマンインターフェイスや福祉などへ．

- アプローチ：自動から人との協調へ

ここでは，以上の流れの中での筆者のアプローチを述べるとともに，今後の課題を考えてみたい．

2 人工知能としてのビジョンの誕生

人工知能の一分野としてのビジョンは，人工知能の創始者の Minsky や McCarthy の影響を受けていた．Minsky は次のように話していた．「大人の知能活動より子供の知能活動の方がコンピュータで実現ことがむずかしい．例えば，積木で遊ぶための知能はとエキスパート知識よりむずかしい．」

ビジョンにおきかえると，局面を限定して詳しく認識するより，局面を限定する方がむずかしい．例えば，シーンの中から拡散反射をする一様な色領域を求めることは，それが与えられたときにその3次元形状を復元するよりむずかしい．

ビジョンをそれまでのパターン認識と比較すると，その主な特徴は

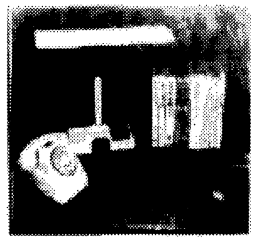
- 3次元の対象を取り上げたので，見え方が一定でない．
- シーンに複数の対象を含むので，セグメンテーションが必要．
- 処理結果がシーンの記述である．

つまり，位置決めされた単体の認識より，セグメンテーション方が困難であるので，それに取り組もうとした．

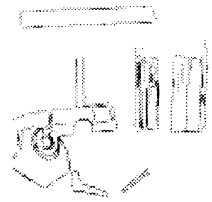
そのアプローチは，特徴（線画）を抽出してから特徴を解釈することであった．これに対し，Minsky は，処理を順番に行なうという階層的な方式は，途中での間違いが増幅されるのでよくないといった．

問題は間違いを犯すことだけでなく，どれだけの特徴を抽出すればよいかがわからないことである．人間は大体の状況を認識してから必要に応じて細部を見る．コンピュータで大体の状況の把握するには，明確な特徴をてがかりにすればよい，そこで，まず明確な特徴を抽出して認識を試み，それで充分でない場合には，さらに詳細な特徴を抽出て認識を行なうという方式が必要であると考えた．

Minsky の指摘の直後の 1971 年に，MIT で1年間滞在する機会を得た．偶然，Binford-Horn の Line-Finder [1] を動かしたら，間違いの多い線画を出力した．これがきっかけで，非階層的な方式の積木認識を行なった [2]．論文の査読で，タイトルとして"Context Sensitive Line Finder..." を勧められた．認識まで行なっているので不満だった．



入力机上シーン



認識なしで得た線

最初に得た線

最終結果

図 1: 机上シーン認識の例

MIT では，積木を置く台が黒いため，明確な特徴を抽出すると，積木と背景の境界が得られるので，提案方式のよさを発揮できていない．電総研に戻ってから，大型プロジェクト「パターン情報処理」の一部として，灰色の机の上にある，より明るいものや暗いものの認識を行なった．最初に最も明確な特徴を抽出し，それを解釈し，必要ならばさらに詳細な特徴を抽出するという繰り返しを行なった [3]．

対象物の線画は曲線を含むので，線画を直

線と楕円で近似する方法，モデルの表現法などにかなりの労力を使った．デモを行なうので，ロバストにしなければならないこともあり，かれこれ5年近くも要した．入力画像全体をメインメモリに蓄えない，コンピュータの入力にカードを使った，処理が遅かったことなども研究の長期化の原因であった．

3 入力情報の拡大

3.1 初期のCVへの適用

初期のCVは，単眼視濃淡画像から3次元シーンを認識することが大きな課題であった．積木の線画の認識の直接の目的は，認識結果に基づいて積木を掴むことであった．

一方，ロボット移動のためには，距離の直接測定を行なっていた．それは，認識した物体の一点に光電管を向けながら，スポット光を走査し，最も明るくなった時のスポット光の方向から3角測量の原理で距離を測る方法であった．

私は，1969年にCVに参入した時，単眼視による積木の認識は原理的にむずかしいので，入力情報を増やせばいいと考えた．

その1つは，シーン全体の距離情報を得ることであった．カメラ全体を光電管とみなし，一度に多数のスポット光を走査するかわりに，平面状の光で走査する方法を考えた．早速同じ研究室の諏訪氏が装置を作り，レンジファインダの試作装置が完成した．おかげで世界初の距離データだけからの積木認識が実現した[4]．

その直後，Stanford大のBinfordのところで，レーザを用いたレンジファインダが作られ，曲面物体の認識が発表された[5]．その後数年間は，距離データを用いた認識は入力装置のあるところに限られていた．

もう1つは，エッジの抽出法である．それがむずかしい理由は，積木の線画抽出には稜をエッジとして検出しなければならないことであった．照明条件によっては稜の明るさのコントラストが充分でないので，エッジが検出できない．また，積木の影の境界でもエッジが検出される．したがって，理想的な線画を得ることは理論的に不可能である．ところが，3種類の異なる光のいずれかの照明で稜の両側の明るさが異なる確率は大きい．また，光の方向が異なれば，影のできる場所が異なる．そこで，独立な4方向から光を順番に当て，それぞれの線画を抽出し，それぞれの線画の論理演算によって完全な線画を得る方法をIJCAI'71に発表した[6]．当時は反響がなかったが，20年程後に，「国際会議に論文が採択されたが，すでにあなたの論文があることがわかった」というメイルをもらった．

3.2 オプティカルフローと距離の統合

追跡の研究はたくさんある．目標物体の明るさや形状が変化し，カメラが動く場合は，背景差分や相関に基づく手法がが使えないので，それ以外の手がかりが必要となる．

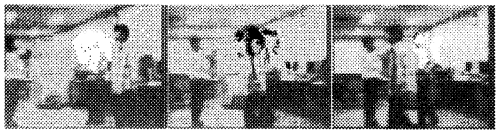
オプティカルフローを使えば，物体の形状や輝度パターンが滑らかに変化する場合にも追跡を行なうことができる．目標物体の周辺に似た速度の物体がある場合には，目標物体を区別できない．そこで，ステレオカメラからの距離情報を用いることができる[7]．

目標物体が隠蔽されてない状態での領域の更新は以下のように行なう．画像上の点$\boldsymbol{p}_i$で観測$\boldsymbol{o}_i = (u_i, v_i, d_i)$が得られた時，その点が目標物体($T$)に属する確率が高い点を目標物体領域とする．目標物体上の点で$\boldsymbol{o}_i$が観測された場合の確率は，

$$P(\boldsymbol{p}_i \in T|\boldsymbol{o}_i) = \frac{P(\boldsymbol{o}_i|\boldsymbol{p}_i \in T)P(\boldsymbol{p}_i \in T)}{P(\boldsymbol{o}_i)}$$

分子の第1項は観測モデルから，第2項は前フレームからの予測から得られる．

目標物体が隠蔽されている場合は，隠蔽という事実を検出し，隠している物体を追跡しながら，隠されている物体が現れるのを待つ．本手法で追跡を行なった結果を図2に示す．

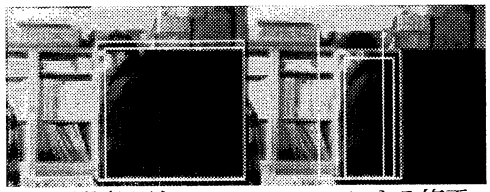


No. 20　　No. 40　　No. 50

図 2: フローと距離による追跡例 (白: 目標物体領域, 黒: 目標物体を隠蔽する物体領域)

3.3 フローと明度一様領域の統合

コントラストがない場所ではオプティカルフローと距離情報が得られないので, 明度が一様な領域を追跡できない. その場合は, オプティカルフローと明度一様領域を組合わせる方法がある.



明度だけ　　フローによる修正

フローだけ　　明度による修正

図 3: フローと明度一様領域による追跡例

目標領域内ではコントラストのある部分でフローが得られ, コントラストのない部分で明度一様領域が得られる. 得られた目標のフローまたは明度一様領域のうち, 少なくともどちらか一方が背景や他の物体の領域と区別できる場合に目標物体を追跡することができる [8]. 16 個の DSP を用いた実時間追跡の例を図 3に示す.

情報統合の研究は, 1 つの情報の処理法の提案でないため, 成果がわかりにくいのが難点である. センサーヒュージョンの論文特集号で論文 [9] を依頼されたのも, 採録論文が少なかったためと聞いている.

4 何を仮定するか

4.1 対象に関する知識があればいいか

CV の理想は, どんなシーンでも認識することである. そのためには, 世界に関する知識が前提となることが指摘されていた. 自然言語や知識工学の分野の方が先に知識を必要とした.

1977 年に, ヒューリスティックな発見の研究で Computer and Thought 賞を受けた Lenat は, その後対象分野が限られていて, 少し分野から外れると全く対応できないことに不満をもち, 80 年代に大規模な知識ベースシステムの構築を始めた. このプロジェクトは CYC とよばれ (Lenat と CYC をキーとしてインターネットで検索可能), 10 年以上続いた. 大規模知識ベースはできたが, それによって知識を利用したすばらしいシステムができたという話しはあまり聞かない.

CV でも, 1970 年代後半から対象に関する知識の必要性が重視されたが, 例えそれがあったとしても, 実データを使った認識には, もっと他にも重要なことがあることがわかった.

それは, 知識を適用するための最小単位の情報を実データからとりだすことである. これはパターンから記号への変換ともいわれ, 難しい課題とされている. これができれば, 後の処理は自然言語処理などと類似するところが多い. 多くの音声認識システムは, ユーザが語彙や文法を限定するか, 対象を新聞記事のように緩やかに限定して求めた単語の出現確率に基づいた認識のいずれかである.

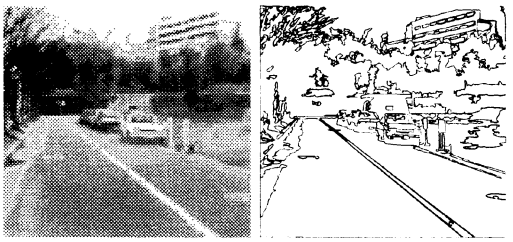
CV では, 拘束をなるべく少なくしてある程度の汎用性を目指すか, 拘束をつけて有用性をねらうかの 2 つのアプローチがある.

4.2 汎用的な知識の利用

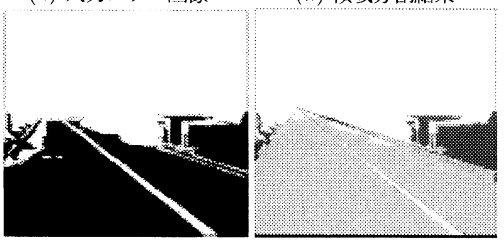
対象シーンをある程度限定して認識を行なう場合, やるべく拘束をゆるやかにしたい. その一つは広く適用できる知識を使うことである.

1988年代末に車の自動運転でのための道路領域認識の研究[10]が盛んであった時，より大局的道路を見つけたり，目的地を認識する研究例[11]を示す．

まず，カラー画像を領域分割しておく（図50）．もちろん，得られた領域は実世界の意味のあるものと対応していると限らない．



(a) 入力カラー画像　(b) 領域分割結果



(c) 道路候補領域　(d) 道路領域（灰色）

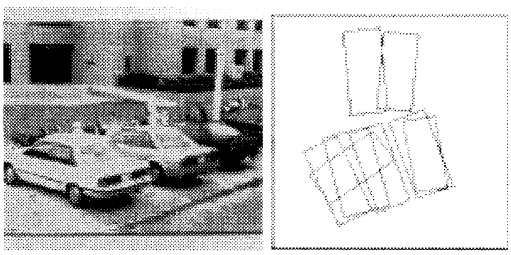
図 4: 道路シーンの認識例の一部

道路領域抽出のため，「道路は画面の下の方にあり，大きく，色がほぼ一様で薄い」という知識を用いて候補を出す（同図(c)）．つぎに，「通行区分線は道路上にあり白か黄色で，実空間で並行である」という知識を用いて，通行区分線を抽出し，「通行区分線のある道路はそれに囲まれている」より，道路を限定することができる同図(d)）．このように，弱い拘束を与える知識を何度も適用することによって，普通の道路，車，木，建物などを認識することができる．しかし，対象が画像上で常識的な性質をもっていない場合はむずかしい．

4.3 確率的な知識による確率的認識

パターン認識理論では確率的決定論が常識である．認識の結果が対象の名前であるから，限定された答に対して確率を求めればよい．CVでは，シーンの記述を作るので，答を限定しにくいので確率の導入を困難であった．それでも，1973年にYakimovsky [12]が確率的解釈を行なって注目されたが，確率的知識を得ることが困難であることと，計算量が非常に多くなるため，その後しばらくは試みられなかった．

我々は，1990年代初めに屋外風景解釈に確率的なアプローチを試みた．入力データは，カラーステレオ画像で，最初にエッジの距離を求め，カラーによる領域分割しておく．図5(a)に示すような複雑な対象では，非常に多くの領域に分割される．



(a) 左のステレオ画像　(b) 車に関する仮説

図 5: 確率的な解釈例の一部

一つの構成要素に一つの領域が対応するとすれば（実際は異なるので，前処理が必要），領域の集合 $(R_1, R_2, \cdots, R_N)$ が，位置や姿勢が決められた物体の集合 $(O_1, O_2, \cdots, O_M)$ として解釈される信頼性はベイズの定理によって，以下のような確率として表すことができる．

$$P(O_1, O_2 \cdots O_M | R_1, R_2, \cdots, R_N) = \frac{P(R_1, \cdots, R_N | O_1, \cdots, O_M) P(O_1, \cdots, O_M)}{P(R_1, \cdots, R_N)}$$

上式は解釈可能な物体のシーン中での位置や向きのすべての組合せに対して計算する必要がある．　組合せを限定しないと計算できないので，画像特徴から可能性のある組合せを求めなければならない．この部分は確率を用いない方法と同じで，その場合は，最も可能性の高いものだけを求めていることになる．車に関する可能性の高い組合せの例を図5（b）に示す[13].

ここで問題となるのは，上式の中の事前確率 $P(O_1,\cdots,O_M)$ である．ありそうな物体の組合せに高い確率を与えるのは尤もであり，人も定性的に知っているが，どのように決めたら納得されるかである．多くの論文（この例も）では，適当なサンプルを用いた学習によって求めているが，それで充分であろうか？

4.4　特徴抽出法が既知でない場合

人は未知の画像を見てもそれなりに解釈できる．その時，どのような特徴抽出を行なっているのであろうか？

例えば，図6を見ると，林の中の大きな木がわかる．木の左の輪郭はわかりやすいが，右ははっきりしない．木であることがわかると，ほぼ左の輪郭と平行という仮定から推定できる．ということは，木であることは右の輪郭を知る前に知っているのであろうか？多分，不明確であるが，右の輪郭もある程度抽出し，木であるという仮説を作ってから検証しているのであろう．それでは，どのようにして輪郭を求めるのかが問題になる．

この問題に取り組んだことがある[14]．人は多様な特徴を多様な解能とパラメータで抽出しているのであろう．この場合も，対象に関するる複数の仮説をもち，複数の特徴抽出法を適用する．その結果の例を図7に示す．右からは幹の相当する部分が見える．さらに，抽出点の密度の濃い部分の領域を求め形から知識を使って仮説を作る．　この方法では，人の脳が並列処理をしている膨大な特徴抽出の組合せを試みなければならいのが問題であり，これ以上先に進むのは困難であった．

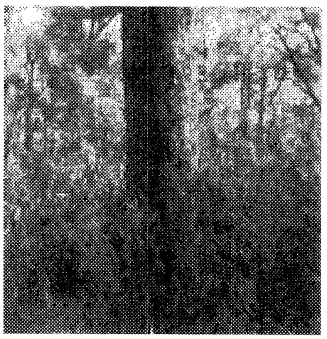


図 6: 屋外画像の例

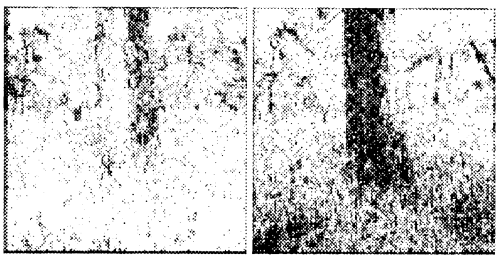


図 7: 幹としての確からしさ

5　インタラクティブビジョン

人のビジョンに迫るCVを研究することは夢があっておもしろいが，なかなか実用にならないことが欠点である．現在，CVはある程度成熟した分野となり，有効な応用が期待されている．新しい応用は，新しい課題を生み，それが理論の発展にもつながるので，挑戦的な応用は研究の源ともいえる．

もう1つの方向として，人との協力がある．例えば，建物，文化財，踊りなどの3Dディジタル化は，人が介入していいものを作ることが目的である．

同様に，福祉への応用でもユーザとの協力を想定できる．5年程前から冷蔵庫から頼まれたものをもってくる研究を開始し，3年前

には2年間の萌芽研究「インタラクチティブビジョン」を立ち上げた．そこでは，完全自動のCVシステムを作るのでなく，ユーザとのインタラクションによって目的を達成することが課題となる[15]．

対話型ロボットでも，対話のための音声認識や言語処理，CVの基本機能を備えていなければならない．それに，ものを取ってくるという課題に依存する機能が加わる．後者に関するCV側の要求は以下である．

1. 頼まれたものの検出を試み，その結果が成功，失敗，不明，複数の候補が見つかる，隠蔽などの結果を得る．

2. 必要な情報を得るためにユーザへ画像と音声を出力する．

3. 間違えた場合の学習と照明条件の影響の学習．

まず，1.では，適切な物体の色や形のモデルを用意しておき，それとの照合を行ない，照合の結果に応じて，状況を判断する．

つぎに，2.では，処理結果を画像で表示するとともに，ユーザに確認をえたり，質問したりする．この時，CVを知らないユーザに，システムの状況をわかってもらうための表示と発話が重要である．この部分はCVの基本機能にも依存する．情報を得る場合も，一度のやりとりか，数回に分けて可能な場合を絞っていくかなどの計画を立てなければならない．どの計画が適切かは，音声認識の誤り確率も考慮しなければならず，この部分に対してはまだ基本方針確立していない．

最後に，3.では，間違えた場合に対話で修正した後に，今後に備えておく．特に，照明条件の変化により，対象の色がどのように変化するかを学習することは自動認識の能力向上に効果的である[16]．

実用化のためには，以上のCVの他に状況に応じたユーザへの発話と音声認識が必要である．既存の音声認識エンジンは新聞の記事のような文からデータベースを作っているので，このような応用には適さない．また，すべて既知の単語を用いるなら，簡単に対処できるが，未知語にも対応しなければならない．現在はこのような応用に耐えるものはないので，既存のエンジンをそれを目的に合わせて拡張するか，あるいはユーザが後処理を行なう必要がある[17]．

6 おわりに

これまで，筆者が行なってきた研究の一つの側面として，人のビジョンへのアプローチ，入力情報の拡大，インタラクティブビジョンを概観し，その課題を述べた．これ以外にも多数の課題がある．筆者の経験したものだけでも

- ビューベイスの手法の限界： 画像の認識に特別な特徴抽出を行なわなくてもいい例として，固有空間法，ニューラルネットなどがある．ところが，対象の位置や大きさが正規化されていないと適用できない．正規化されていないと，正規化のために特徴抽出が必要となる．例えば顔認識では，眼や鼻などの位置を用いて正規化することが多いが，顔の方向によっては，特徴の位置決めが困難（できたら当日例を示す）．さらに，照明の方向によっては眼の位置の検出に誤差がでる．本当は，特徴を正確に抽出しなければならないのではないか．

- センサーヒュージョンによるシーンの記述： もし多くのセンサー情報があるとシーンの記述ができるか？心理学の実験によると（私も体験），屋内シーンを左右を反対の偏光眼鏡でステレオ視すると距離は逆転するが，それはほとんど感じない．シーンを認識しているので距離の手がかりは使っていない．距離は，輪郭の連続性，大きさなどからもわかるが，

いずれも認識をしてから使える情報である．セグメンテーションと認識は切り離せないところが，記号処理と異なりおもしろいところであろう．

参考文献

[1] B.K.P. Horn: The Binford-Horn Line-Finder, AI Memo 286 (1971) あるいは http://www.ai.mit.edu/people/bkph/AIM/AIM-285-OPT.pdf

[2] Y. Shirai: A Context Sensitive Line Finder for Recognition of Polyhedra, Artificial Intelligence, Vol.4, No.2, pp.95-119 (1973)

[3] 白井: 濃淡画像から複雑物体を認識する一手法, 情報処理, Vol.17, No.7, pp.611-617 (1976)

[4] Y. Shirai and M. Suwa: Recognition of Polyhedrons with a Range Finder, Proc. 2nd IJCAI'71, London, pp.80-87 (1971)

[5] G.J. Agin and T.O. Binford: Computer Description of Curved Objects, Proc. 3rd IJCAI, pp.629-640, (1973)

[6] Y. Shirai and S. Tsuji: Extraction of the Line Drawing of 3-Dimensional Objects by Sequential Illumination from Several Directions, Proc. 2nd IJCAI, pp.71-89 (1971)

[7] 岡田他: オプティカルフローと距離情報に基づく動物体追跡, 電子情報通信学会論文誌, Vol. J80-D-II, No. 6, pp./ 1530–1538 (1997)

[8] 山根他：オプティカルフローと明度一様領域を統合した人間の実時間追跡, ロボット学会誌, Vol. 18, No. 4, pp./ 521–528 (2000)

[9] Y. Shirai et al: Robust Visual Tracking by Integrating Various Cues, IEICE Trans. on Inf. & Syst., vol.E81-D, no.9, pp.951-958 (1998)

[10] C. Thorpe, et al: Vision and Navigation for the Carnegie-Mellon Navlab, IEEE Trans., Vol. PAMI-10, No. 3, pp.361-372 (1988)

[11] Hirata, S et al: Scene Interpretation Using 3-D Information Extracted from Monocular Color Images, Proc. IEEE/RSJ Int. Conf. on Intelligent Robotics and Systems, pp.1603-1610 (1992)

[12] Y. Yakimovsky and J.A. Feldman: A Semantics-Based Decision Theory Region Analyzer, Proc. 3rd IJCAI, pp.580-588 (1973)

[13] 谷口他: 認識の信頼性を考慮した屋外シーンの解釈, 電子情報通信学会論文誌, Vol.J80-D-2, No.6, pp.1493-1501 (1997)

[14] M. Hild et al: Scene Interpretation with Multi-Parameter Default Models and Qualitative Constraints, IEICE Trans. Inf. & Syst., Vol.E76-D, No.12, pp.1510-1520 (1993)

[15] Y. Makihara, et al: Object Recognition Supported by User Interaction for Service Robots, Proc. ICPR, pp. 561-564 (2002)

[16] 槇原他: 対話を用いた物体認識のための照明変化への適応, 電子情報通信学会論文誌 D-II, No.2, pp.629-638 (2004)

[17] 滝澤, 槇原他: サービスロボットのための対話システム, システム制御情報学会誌, Vol. 16, No. 4, pp. 174-182 (2003)