Vol. 23 No. 2 情報処理学会論文誌 Mar. 1982

# 言語・画像を利用した行動の解釈(1)†

## ——発話指示による対象の同定——

安部憲広†† 曽我巌哉††† 辻 三郎††

簡単な画像とそれを説明する英文を与え，その両者に含まれる情報を利用することによって，画像または言語的知識だけでは，処理が繁雑であったり，解釈が一意に定まらない問題が，解決可能なことを示す．このような研究を行うには，従来独立に研究されてきた言語と画像処理とを結合することが必要であり，本研究はそのような研究の本格的試みの最初のものである．研究遂行のためには，画面内に出現している対象物体を同定し，物体相互の関係を抽出するプログラムが必要であるが，同時に説明文を解析するプログラムと，それら両者の結果を利用して，それぞれの処理を進展させる手続きも必要である．

本論文では，これらの多くの処理手順のうち，特に画面内に存在すると説明文で述べられた物体を同定する過程について述べる．すなわち，対象同定に必要となるモデルの宣言的記述，部分的照合機能を有するモデル同定の手法について述べる．この対象の同定と，対象間の関係記述は，主要な対象，関係はあらかじめ抽出されるが，画像に含まれる雑多な関係記述は，座標のまま，または原画のまま残される．その理由およびこのようなデータを管理するフレーム構造についても述べる．

## 1. はじめに

近年人工知能と呼ばれる分野に含まれる研究が盛んであり，その一分野を占める物語理解の研究も数多くなされている．この研究は現在わが国においては，あまり研究がなされていないが，世界的にはさまざまな側面から研究が行われている．それらは大別すると，

① 言語処理や表現の問題として物語を把握するもの，

② 画像情報を利用して行動を推論するもの，

となる．

前者は Schank らを中心とするグループによって研究が遂行されており，よく知られた script[1] の他に，繰り返すゴールの概念を用いるもの[2]，また人間の性向を参照する[3] 物語の理解が試みられている．それらに共通した考え方は，他の人工知能の分野における問題解決の基本的戦略である，ゴールとそのためのプラン作成という基本的思考方法を用いている．

後者としては，我々の研究室で始められた動画中の行動の推論がある[4]．しかし，行動様式だけから具体的行動内容を知ることは，非常に単純な行動様式は別にして，仲々困難なことがわかった．我々が動画を観察して得るのと同じ筋に，プログラムが到達するためには，非常に多くの常識が必要であると同時に，行動推論に導びく関係抽出のために，視界に存在する雑多な事物のどれに着目すれば良いかという点の解明が必要であるが，現時点では十分には対処できていない．

この何に注目すべきかという問題は，ゴールの概念と並んで，人工知能の大問題の1つであり，Schank は interesting をその基本要因として挙げている[5]が，現在の研究水準でこれを，アドホックでない方法で，プログラム化することは，非常に困難である．何らかの注視点を与える方策が必要である．

視界に出現する事物の中で，何に注目し，どんな関係を抽出させるかを示唆するものとして，本研究では発話による指示を考えている．発話指示された物体，およびそれと密接に関連する事物を着目すべき対象と考え，以後それらの事物間の関係を追跡することにより，筋を抽出する．すなわち，解説（以後ナレーションと呼ぶ）付きの簡単な絵を提示して，その絵（以後，画像と呼ぶ）に出現する物体間の関係と発話の関係を抽出させることにより，粗筋の把握を試みる．

本研究は，物語理解以外にも，後述するような目的を有しているが，現在システムの取り扱い得る物語は簡単なものであり，我々がそれを見た場合，実はナレーションがなくても大筋は理解できる．これは我々の持つ豊富な知識によるものである．しかし幼児の場

---

† Interpretation of Act Using both Language and Image Data —Identification of Objects through Utterance by NORIHIRO ABE (Department of Control Engineering, Faculty of Engineering Science, Osaka University), ITSUYA SOGA (Mitsubishi Co. Ltd.) and SABURO TSUJI (Department of Control Engineering, Faculty of Engineering Science, Osaka University).

†† 大阪大学基礎工学部制御工学科

††† 三菱電機(株)

合には，話とともにそれを表現する絵がないと，話はよく理解できない．言語理解の未熟な幼児では，実際的なイメージを与えることによって始めて，全体的構成をつかむことが可能となる．すなわち，言語を理解するということは，内的表現あるいは，それを構成して得られるイメージと言葉とを結び付ける操作だと考えることができる．そうした関連付けが行われると，Waltz が指摘しているように[6]，言語レベルだけでは推論に手間のかかる操作も，簡単になる場合がある．しかし Waltz は，デフォールトな知識の援用が有益な結果をもたらし得ることを述べたに過ぎず，現実に眼前の情景をどうするかについては，全く実験していない．

以上の観点から，言語情報と画像情報を活用した簡単な筋の把握を試みた．なお，本論文では，本システムの考察目標，および全体的な処理の流れを概説した後，特に画像識別の部分に的を絞って解説する．言語処理部，筋の抽出，それに基づく質問応答を含む実際の処理結果例，および本システムの改良すべき点，将来の展望などは，紙面の都合上本稿では記述不能なため，別稿で詳述する．なお，プログラムはすべての処理が，FLISP[16] で記述されている．

## 2. 研究目的

物語理解を目的として本研究が動機づけられているが，方法論的には次の3つと密接な関連を持っている．

① 言語・画像間干渉による問題解決
② 簡潔な記述とその利用
③ 画像に対する言語的アクセスの実験

それぞれについて解説する．

① システムは，画像処理部と言語処理部に二分されるが，両者は個々に独立に動作しているのではなく，互いに処理結果を参照，あるいはサブプログラムとして他方を活用している．これは人間の通常の振舞と以下の点でよく合致している．

■ 人間があるシーンを見た時，確かにそのシーン全体の情報は感覚器に入力されているが，特定の対象以外の物については，その特徴あるいはその存在すらも認識されていない場合がある．全視覚情報のうちで，興味ある対象に関してのみしかるべき認識処理が施されて，高次のデータとされるが，その他は意識にのぼらない低次データとして潜在しているものと思われる．計算機を用いて対象を探索する際にも，興味ある対象のみを，その存在範囲を限定して探索する必要がある．本研究ではそれを言語情報——発話による指示——により行う．

■ 人間が言語情報によって，複雑な位置関係を把握するには，視覚情報が不可欠である．たとえば，人に道を尋ねた場合，言葉だけで説明されるより，図を同時に示された方が理解しやすい．計算機においても，特に複数の位置前置句の係り受け解析は，言語処理の過程で画像を参照することによって，より容易になる．

■ 多義語の意味の決定，あるいは名詞句の指す対象の同定を行う際には，その時点のコンテキスト及びいわゆる知識が必要だが，その中には視覚情報が含まれる．たとえば，He takes an apple in the box. に対して，“take” は「食べる」なのか「取る」なのか，また “an apple” は，どのりんごを表わしているのかは画像情報を活用すると，より容易になる．

② これは上述の「注目物体」と深く関連している．どの物体に着目するかが明確になれば，必要な記述の量も軽減される．それとともに，それらの事物の関係記述も簡潔化される必要がある．簡単な例を挙げよう．2, 3の岩と，数本の木を含む背景内で，物語の主人公がある木の方へ歩いたとしよう．我々は他の木木や岩と主人公の関係は重視せず，目的の木との関係にのみ着目するのが普通である．しかし客観的事実としては，主人公と他の物体との関係も変化しているのである．これらをすべて記述すると，話の展開とともに多くの物が出現してくれば，記述量が膨大になる．その多くを注目すべき物体と見なさぬことにより，記述を略したとしても，注目すべき物が話の進行とともに増加してくるならば，急速に記述は増大する．この時，どの関係を抽出すべきかは，アドホックな方法を別にして，適切な方法は発見できない．発話指示された行為を中心とする記述以外のすべては，画像から得た座標データもしくは原画像のまま残しておき，より詳細な情報は必要に応じて，これらの情報源にアクセスして求めればよい．

③ 上述した事柄は，画像理解へのアプローチで採られる方法論をも提供している．すなわち，質問された事が要求するデータを適宜探索して，質問を処理するために必要な操作を行えばよい．現在，画像処理は処理専用のコマンドを用いて行われているが，コマンドで実現されるシステムとの会話は非常に制限されたものとなっている．処理途中で，モデルの定義を行っ

たり，部分図形の処理などを指示することが十分できない．ユーザの指示する関係を，その時点で得た知識と，座標データもしくは原画を組み合わせて，抽出する手続きが不可欠である．

本システムの質問応答では，まさにこの手法が用いられている．現在のところ，モデルなどの定義を言語指示によって修正する機能は実現されていないが，本システムはこの問題への拡張性を有している．

## 3. システムの概略

本研究の対象例の一部分を**図1**に示す．簡単なカラーの線画と，それを説明するナレーションを入力し，ナレーションに出現する物体，およびこれに密接な関連を持つ物体を画像から抽出すると同時に，言語表現に解釈の可能性が複数個ある時，画像情報を利用して，これを解決する．そして，言語表現に関連する記述と，画像情報に関連するデータを保持する世界のモデルを構築した後，これらを参照して質問に答える．**図2**にシステムの処理手順の概要を図示する．ただ，現時点では，システムの取り扱う画像に関して，次の仮定を設けている．

（1） 線画は，ナレーションの記述する行為が完了した時点での状態を表現しているものとする．「りんごを食べた」に対しては，食べてしまった時点を図は表現していると考える．言語処理の研究では，このアスペクトに関してもさまざまな考察が行われていて[8]，発話内容からそれが行為の初期，中期，完了のいずれを指しているか，また行為は継続中か，繰り返されているかなどがある程度分析可能である．これに対応させて，1つの行為をこれらの相に分け，各相での典型的パターンをモデルとして与えることも可能であるが，本システムでは，そのうちの完了時点のみを考察対象としている．

（2） 線画の検索は，主として対象の各部分の位置

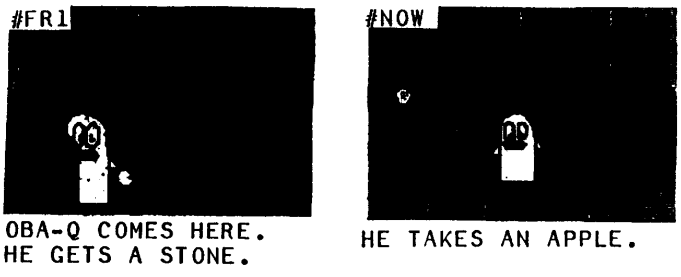



**図1** 入力画像とナレーションの一部
Fig. 1 A portion of input figures and narration

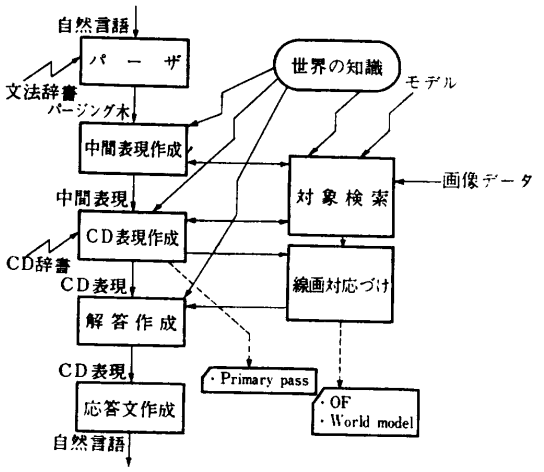



**図2** 解析の流れ
Fig. 2 A flow of analysis of the story.

関係および色によって行い，対象の回転はどの方向にも起きないものとする．最初の仮定は，対象の同定が言語指示によって起動されるという点から，肯定される．すなわち対象の同定は，モデル駆動(model-driven)であり，データ駆動 (data-driven) ではない．言語指示された対象は必ず視界中に存在しているため，可変テンプレート・マッチ[9]のような，厳密な線分の照合を考える必要はない．これは我々が，新聞などの4コマ漫画を見る場合を考えても明らかである．別のコマに出現する絵は，厳密な意味では以前の絵とは一致していない．多分同一の人物，物が引き続き現われるだろうという仮説を認めているから，粗い照合で理解が進むと考えられる．

後者の仮定は，取り扱っている画像が2次元的なものである点と，処理手続の簡単化のためである．主人公が「逆立ち」をしたり「寝ころぶ」という場合は，モデルの 180°，90° 回転が必要であるが，現在は考慮していない．

なお，モデルは宣言的な形をしていて，文献 4) のような手続き的な形を採っていないし，かつ文献 10)，11) のような方法とも異なっている．これは，前述したように，将来モデルの学習を試みる場合に適しているといえる．

## 4. 知識の表現

一般に，人間が日常的生活を営む上には，内包的な事項を持つことが不可欠である．同様に，本システムでは，世界の知識 (world knowledge)，画像モデル，言語処理用の辞書

をあらかじめ知識として有している．そして，これらを活用して，因果鎖で結合された内部表現，位置的な世界モデル，および各対象に関する情報を格納した対象フレーム (object frame; OF) を作成し，これらを使って質問に答える．本稿では，対象物の同定，およびその属性決定に要求される知識，すなわち画像操作に必要な知識を詳述する．(他は文献 15) で述べる)．

### 4.1 世界の知識

人間が物語を理解する場合，その背景として，物語が展開される世界に関する暗黙的な了解事項をすでに所持していると考えられる．本システムではそのような知識を一般／特殊 (generic/specific) の関係で階層化されたセマンティック・ネットワークに組織した．ネットワークは対象，対象の属性，概念等を表現するノード (node) と，ノード相互を意味的に連結するリンク (link) から成っている．ノードはフレーム構造をとっていて，そのノードが記述している対象の持つ一般的な性質，属性を表わすスロット (slot)，および画像検索用のモデルを含んでいる．またネットワークの最末端のノードは，個々の対象に関する記述を持つ OF (object frame) になっている．一方，リンクは 1 組のラベルと値の対を持っていて，各ノード間の関連を表現している．**図 3** にネットワークの一部を図示してある．当然のことながら，文献 12) と同様，情報の継承が可能である．

### 4.2 OF (object frame)

図 3 に示したように，最も具体的な個々の対象物に関する事実を格納するフレームである．そのスロットは，物体の存在する位置，性質，関連する行為，モデル記述と画像データとの対応などから成る．属性の非継承[12]は，たとえば，りんごは普通は赤いが，あるりんごが黄色なら，その OF に黄色と書くことによって，実現できる．

モデルと実画像との対応は ＊PLIST というスロットの下に記述されていて，モデルに記述された部分領域名を PR$n$，対応する画像内での領域番号を R$n$ とすると，＊PLIST ((PR1. R1)… …(PR$n$. R$n$)) という構造を持っている (詳細は図 10 参照)．

前述したように，物体間の関係記述に伴う，

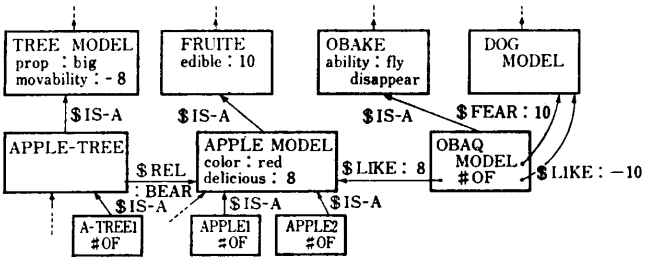



**図 3** 世界の知識の例

Fig. 3 An example of the world knowledge.

対象の記述

```
(*PICT object-name threshold-score)
  <relation-statement-1>
        ...
  <relation-statement-n>
```

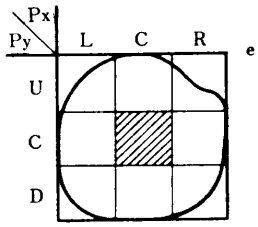
部分領域の記述

```
(*SUBR part-name threshold-score)
  <color-statement>
  <relation-statement-1>
        ...
  <relation-statement-n>


<color-statement>
  := (*COL color score)


<relation-statement>
  := (part-name relation position score)
   or (SUBP part-name score)


relation ┬ IN ── 内部
         ├ OUT ── 外部
         ├ CIN ── 接触して内部
         └ COUT ── 接触して外部
position : ((Px) Py)
```




ex. 斜線部分
((C) C)

枝領域の記述

```
(*SUBB part-name threshold-score)
  <color-statement>
  (<branch-structure> score)


<branch-structure>
 := [bname-1 structure-1 bname-2 structure-2
     ... bname-n structure-n]
 or NIL ：終端枝の場合


bname := branch-name
      or part-name


structure := <branch-structure(next level)>
          or connected-node-name
```

**図 4** モデル記述の仕様

Fig 4 A specification for the model description

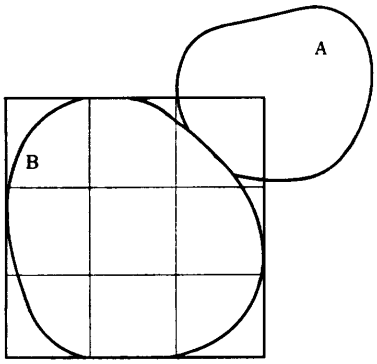



図 5 領域AとBの関係
Fig. 5 A relation between two regions A B

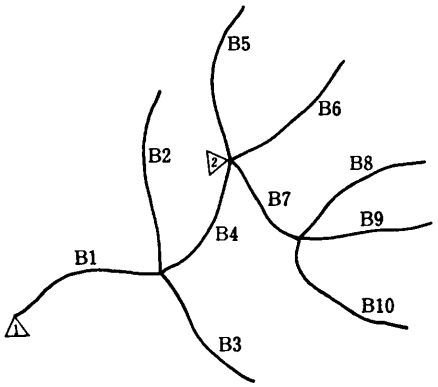



図 6 枝領域の例
Fig. 6 An example of a branch structure.

組み合せの増大を防ぐために，発話指示されない限り，位置関係は座標のまま表現する．

### 4.3 対象のモデル

対象モデルを設定する際の仮定として，

（i） 線画は若干の奥行きを持つ2次元的図形，

（ii） モデルの詳細な形状は考えず，対象の各部分間の位置関係および色によって記述する，の2点を設けた．

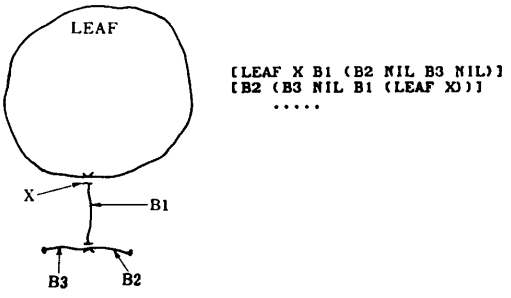



図 7 ステック・フィギアとその記述
Fig. 7 A stick figure and its description.

前者についてであるが，正確な距離情報がなくても，物体が規準値より小さかったり，また以前に出現した時の大きさに比して小さければ，遠くにあると考えるのが常識である．そこで，遠方，近方，普通の距離に対応した奥行き情報を設定した．

後者の仮定は次の理由による．人間は複雑な対象をそれを構成する単純な部分の複合体として捉えることが多い．しかも，その各部の形状が多少変化していても，あるいは一部が全く歪であったり，見えなくても，その物体を認識できる．このような柔軟性をモデルに持たせるために，(ii)の仮定がある．

モデルの形式を図4に示す．対象はまず，*PICT 文によって宣言され，個々の部分領域 (*SUBR, *SUBB) に分割され，階層的に記述される．各部分領域相互の関係は IN (内部)，OUT (外部)，CIN (接して内部)，COUT (接して外部) で表わされ，一方の部分に対する他方の位置は，図4の例のように，領域に外接する長方形を九分割し，構成部分体が，そのどの小長方形に入いる(接する)かにより表現する．したがって図5の例では，領域AとBの位置関係は (A COUT B ((R) U)) と記述される．また部分領域の中には，枝領域と呼ばれるものもある．これは線分のみから形成される領域である．例を図6に示す*．このような構造に対しては，

⚠から記述を始めると，

(B1 (B2 NIL B3 NIL B4 (B5 NIL B6 NIL B7
(B8 NIL B9 NIL B10 NIL)))) となり，

⚠から記述すると，

(B4 (B1 NIL B2 NIL B3 NIL) B5 NIL B6 NIL B7
(B8 NIL B9 NIL B10 NIL)) となって，

表現は一意的には定まらぬが，一方から他方へは簡単に変換が可能であり，その照合に問題は生じない．枝領域と，通常の領域とが接した場合の枝構造の記述例を図7に示しておく．したがって，本モデルは stick figure[13] も表現することができる．

実際のモデルの例を図8に示す．*PICT 文で DESK という対象のモデルであることが宣言され，それが DSK という部分から構成されていることが①で示されている．そして DSK は②の示すように，色が茶で，FOOT 1，FOOT 2 という部分領域と，右下，左下の位置で COUT の関係にあることなどが，階層的に記述されている．各ステートメントの最後には，スコアが記述されていて，これは部分的照合 (partial

* 記述は，図4の表現方法に基づく．

```
①[(*PICT DESK 25)
      (SUBP DSK 30)

②[(*SUBR DSK 25)
      (*COL BROWN 10)
      (FOOT1 COUT ((R) D) 10)
      (FOOT2 COUT ((L) D) 10)

  (*SUBR FOOT1 10)
      (*COL BROWN 10)

  (*SUBR FOOT2 10)
      (*COL BROWN 10)
```

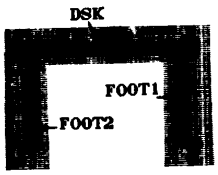



図 8　DESK のモデル

Fig. 8　A model for a desk.

match) に使われる．これは，前述した柔軟性を持たせるためにある．

## 5. 対象の同定

理解を容易にするため，簡単な仮想のモデルを用いて，対象の同定方法を解説する．検証する部分は，

```
( (*SUBR A thsc)                    :〈S0〉
  (*COL B scA1)                     :〈S1〉
  (C rel-2 pos-2 scA2)              :〈S2〉
  (D rel-3 pos-3 scA3) )            :〈S3〉
```

という構造を持っているとする．

### 5.1 モデルの評価

**5.2** で述べる減点法によって，S1, S2, S3 のそれぞれに対して，減点 **xA** 1, **xA** 2, **xA** 3 がなされたとする．その結果 $\sum_i (\mathrm{scA}i - \mathrm{xA}i) \geq \mathrm{thsc}$ が成立すれば，領域**A**は検証されたという．これと同じ操作を1つの対象に行って，対象を同定する．

### 5.2 減 点 法

（i） 色情報〈S1〉について：

色がBに等しいなら xA 1=0, そうでなければ xA 1 =scA 1

（ii） 部分関係〈S2〉〈S3〉について：

〈I〉 領域**C**を検証して，発見できれば

$$y_1 = \frac{\mathrm{scA}\,2 \sum_i \mathrm{xC}i^*}{\sum_i \mathrm{scC}i} \qquad (1)$$

とし，発見できなければ，$y_1 = \mathrm{scA}\,2$ とする．

〈II〉 **A**と**C**の関係が，接しているか否かで異なる時，$y_2 = \mathrm{scA}\,2/4$ とし，全く異なる時 $y_2 = \mathrm{scA}\,2$ とし，満足されている時は，$y_2 = 0$ とする．

〈III〉 **A**に対する**C**の位置について，

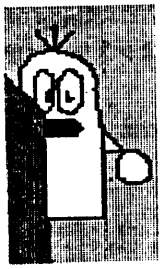
UとC，DとC，LとC，RとCのいずれか1つの組み合せで異なるなら，$y_3 = \mathrm{scA}\,2/4$ とし，一致するなら $y_3 = 0$，その他のは時 $y_3 = \mathrm{scA}\,2/2$ とする．

そして〈I〉〜〈III〉により，$\mathrm{xA}\,2 = y_1 + y_2 + y_3$ とする．

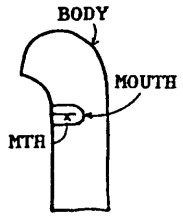
そして**A**と**D**についても同様にして xA 3 をもとめる．

### 5.3 照 合 例

図1の最初のフレームに登場するオバQの照合を例示する**．線画の一部，モデルを**図9**に示す．まず(a)を見るとわかるように，HAND 2 が見えないので，**5.2** の〈I〉によって，①の部分で10点が減じられる．次に②の部分で，BODY に対する MOUTH の関係が CIN になっているため，〈II〉によって，10/4 点が減点される．さらに，MOUTH を評価した時，③の部分で10/4点減点(MOUTH と MTH はモデルでは IN だが，(a)では(b)に示したように CIN になっている）されている影響を受けて，**5.2** の〈I〉の式（1）を使って，$1/4 \times 10 \times 10/(10+8)$ が減点される．その結果 BODY のスコアは 56.7 点となり，thsc =50 を超すので，BODY が検証される．もし，スコアが 50 に到達しない場合は，一応検証失敗として，別の物体を候補として，検証を試みる．そしてすべてが失敗した時は，一番スコアの高い物をオバQとして，残る関係――たとえば，図1ならオバQが石を持っているというような関係――に矛盾がなければ，成



(a)




(b)

```
(*SUBR BODY 50)
      (*COL WHITE 10)
      (HAND1 COUT ((R) C) 10)
      (HAND2 COUT ((L) C) 10) ①
      (HAIR COUT ((C) U) 10)
      (EYE1 IN ((R) U) 10)
      (EYE2 IN ((L) U) 10)
      (MOUTH IN ((C) C) 10)②
       .....

(*SUBR MOUTH 15)
      (*COL PINK 8)
      (MTH IN ((C) C) 10)③
      .....
```

(c)

図 9　オバQのモデルの一部

Fig. 9　A portion of model specification for OBAQ.

---

* C も A と同様な形式で記述されているので，A と同様な方法で xC 1, xC 2, … をもとめる．

** 完全なモデルではない．オバQには毛が3本ある．図10の OF の例の H1, H2, H3 がそれである．

功とする．それが再び失敗すれば，次点の物を選んで同様な処理を行う．(実際例では，高々次点止まりであった)．

### 5.4 対照探索範囲の限定

発話指示された物体を画像内で探索する時，その対象に関連した知識や脈絡を利用すれば，その存在域を限定することが可能となり，照合を行う候補の数を減ずることができる場合がある．

(1) 対象の属性に帰因する束縛

机やいすのような物体は，通常は床または地面の上にある場合が多い．また，りんごのような果物は，もしその情景に木があれば，木になっている可能性が高い（もし，「りんごがテーブルの上にある」というようなことが既知なら(2)の方法を用いる)．このように，ある物体がそれよりも容易に発見し得る物と密接な関係を持つ可能性が強い時は，その物の属性に，この物体を書いておき，利用する．ただ，実際には，この束縛はあまり多くは利用されていない．なぜならば，本研究の対象画像は，前述したように，正確な位置情報を有していないため，地面や床の位置が正しく推測できないからである．地面や床は，単純に画像の下辺であるとして，机やいすを探索するにすぎない．したがって，図1の#NOWのような家や木は，ただちには発見できない．#FRIでは，オバQは地面から迅速に探索される．しかし，#NOWで木が発見されると，りんごを発見するのは容易である．木のなかでりんごを捜すことにより，目的は果たされる．図1の#NOWの場合，ナレーションは明示的には木の存在を述べていないが，りんごが発話指示されていることから，木の存在が検証される．そしてそれが逆に#FRIでと木りんごの探索を起動する．岩は，全く言及されていないため，抽出されずに画像のまま残される．

(2) 支持関係等による拘束

オバQが発見された時，#FRIのナレーションによって，GETが「物を取る」を意味するなら，手が石に接触しているはずである．こうして石が発見され，同時にGETの意義が確定する．また，例には出現しないが，オバQがりんごをどこかに置いたとしよう．その時，その行為が明示的に言及されなかったとしても，りんごの位置が不自然な位置にあれば，それを支持する台のような物体が存在するか否かが調べられる．「台に置いた」と明示的に言われれば，当然その関係から，関与する対象の探索が試みられることになる．

(3) 特別な情報のない時

以前の画面に出現した動かない物体が，現在の画面にも出現している時，それらの画面の座標の関係が既知であれば，それを利用して探索する．そうでなければ全探索に移る．

以上述べた手がかり以外にも，対象の存在域を推測させ得る情報は沢山ある．主人公が「右へ歩いた」と言えば，多分次の画面でも右の方に居るだろうと思われる．しかし，取り扱かっている画像は連続的なものでなく，自由な視点の切換えがあるので，このような情報は利用できない．

## 6. 結　果

図1の#FRIの画像から，オバQ，木を抽出し，それを，そのOFに記した結果の一部を**図10**に示す．図中の*ACTは関連する記述や行為などを指すポインタを示している．また，正確なものではないが，出現物体の奥行き方向の情報も抽出されている．ところで，図1を見るとすぐわかるように，木を仲介役として#FRIと#NOWとは，位置的な関連性が求められなければならない．それらは別稿[15]で詳述する．#FRI，#NOWを表わしている座標系の関係を算出すれば，画面の相互関係，ひいては行為の関連が導びかれる．

## 7. 結　論

簡単な線画と，それを説明するナレーションを用いて，画像中から発話指示された物体を抽出する方法について報告した．しかし本システムに関して，次の点が問題となる．

本論文の提案した方法には，従来の人工知能システ

```
>OBAQ
(*ACT (1 2) *PLIST ((BODY.9) (HAND1.11) (HAND2.??) (HAIR.BR7)
((H3 H2 H1) 32 31 30) (EYE1.12) (BEYE1.13) (EYE2.14) (BEYE2.15)
(MOUTH.16) (MTH.BR8) (((MTH1)34))) *WHERE (65 139) *SIZE (41 74)
*3D (+.655142E+02 +.139291E+03 +.100000E+01))

>TREE1
(*POS (3 4 5 6) *PLIST ((LEAF.2) (TRUNK.1)) *WHERE (151 59) *SIZE
(87 159) *3D (+.130789E+03 +.518121E+02 +.862166E+00))
```

図 10 オバQとTREE1のOFの例

Fig. 10 An example of object frame for OBAQ and TREE1.

ムが必要としていた世界の知識に加えて，図形に関する知識が必要である．そのため，さまざまな状況に対処するためには，必要な知識の総量は尨大なものになる可能性がある．しかし対象とする状況を制限せぬ限り，出現する可能性のある物体の知識は与えておく以外に方法はない．ただその場合，セマンティックネットにより知識を組織化すると，ポインタの関連により，全知識をメモリに収容しなければならない．現システムはこの欠陥を有している．しかし，我々はこの問題の一部に対処する研究も行っている．登場する可能性のある対象の知識を記述する必要性を回避することはできないが，対象となる情景を推測して，必要な知識の一部をメモリにのせることは可能である．それには，スクリプトの考え方を使う．Schank, Abelson はスクリプトを起動する条件として「レストランへ行く」といったような explicit な表現を key としたが，我々はもっとあいまいな語や句から数個〜数十個のスクリプトを引出し，その内から最適のスクリプトを選出する研究を行っている[17]．この結果と，個々の対象のフレーム構造をメモリに読み込む FLISP 関数を使うことにより，必要な知識の取出しが可能であり，本システムへこの方式を採り入れることができる．

次の問題点は，対象の同定がさまざまな状況下で，本論文の主張するほど首尾よく運ぶかという点である．たとえば「木になっているりんご」「皿に盛られたりんご」「虫の食ったりんご」等が同定できるかということである．この点に関しては，本稿の手法は「りんご」の同定に，その宣言的知識と一般的パターンマッチャしか使っていない点に注目されたい．「りんご」が「木」にあるか，「盛られているか」により同定方法は変化しない．「りんご」の存在域を推測するために，「木」や「皿」や「台」を手がかりにするだけである．もちろんそうすると，「5個盛られている」時，正確に5個同定できぬ場合もある．しかし本稿の考え方は，ナレーションに説明された状況に一致するものがない場合，その状況下で他と矛盾しない類似の状況が存在すれば，それをもって同定したとするものである．したがって，上記の問題のほとんどはシステムを惑わせることはない．ただ，類似の状況が複数個ある時，どれを選べば話の筋全体として最も都合がよいかという点では，本システムは不十分であり，改良の必要がある．

問題があるのは「虫の食ったりんご」である．「りんご」は同定できるが「虫が食っている」ことは検出不能である．これは図形のモデルに正確な形の情報が入ってないためである．「腕を曲げる」といったことも同定できない．この点は強化が必要である．しかしその強化によっても「虫食い」の検出は，「りんごの虫食い」に対応するモデルがあれば別だが，やはり困難である．「虫食いりんご」のモデルを用意するのは，モデルの数の増大を招くのみであり，適切ではない．「ある物」に「虫が食う」とどのような気質的変化を生じるかによって，「虫の食ったりんご」「虫の食った葉」などのモデルの合成を行うべきだと考えるが，具体的解決法を現時点で提案することはできない．この問題は Minsky のいった寄生的なフレームによるフレームの合成問題[18]と同種のものである．

以上の問題は，本稿のようなシステムがより現実的状況に適用されるために避けて通ることのできぬ点である．今後，このような点の解決に努力しなければならないと考えている．

## 参考文献


1) Schank, R. C., Abelson, R. P.: Scripts, Plans and Knowledge, IJCAI 4, pp. 151-157 (1975).
2) Wilensky, R.: Why John Married Mary: Understanding Stories Involving Recurring Goals, Cognitive Science, Vol. 2, pp. 235-267 (1978).
3) Carbonell, J.: Towards a Process Model of Human Personality Traits, Artificial Intelligence, Vol. 15, pp. 49-74 (1980).
4) Tsuji, S., Kuroda, S. and Morizono, A.: Understanding of Simple Cartoon Film, IJCAI 5, pp. 609-610 (1977).
5) Schank, R.: Controlling Inference, Artificial Intelligence, Vol. 12, pp. 273-297 (1979).
6) Waltz, D., Boggess, L.: Visual Analog Representations for Natural Language Understanding, IJCAI 6, pp. 926-934 (1979).
7) Charniak, E.: Toward a Model of Children's story Comprehension, AITR-2666, MIT (1972).
8) Steedman, M. J.: Verbs, Time and Modality, Cognitive Science, Vol. 1, pp. 216-234 (1977).
9) Tsuji, S., Osada, M. and Yachida, M.: Three Dimensional Movement Analysis of Dynamic Images, IJCAI 6, pp. 896-901 (1979).
10) 岡田，田町：図形の意味解釈とその自然語記述—要素的図形認識と構造分析，信学論，Vol. J 59-D, pp. 323-330 (1976).
11) 岡田，田町：図形の意味解釈とその自然語記述—意味分析— 信学論， Vol. J 59-D, pp. 331-338 (1976).
12) 田中：日本語の意味構造を抽出するシステム EXPLUS について，信学論， Vol. J 61-D, pp.

549-556 (1978).
13) Herman, M.: A Computer Program for Generating the Semantics of Human Stick Figures, University of Maryland, TR-729 (1979).
14) 曽我，安部，辻：言語，画像間相互参照によるプロットの把握，信学研資，AL 80-85 (1981).
15) 安部，曽我，辻：言語画像を利用した行動の解釈(2)—粗筋抽出と質問応答—，情報処理学会論文誌，Vol. 23, No. 2, pp. 133-141 (1982).
16) 安部，東出，辻：画像処理関数の組込みを中心とする FLISP の改善について，信学論，Vol. J 64-D, pp. 78-79 (1981).
17) 小西：スクリプトを用いた談話理解，大阪大学修士論文 (1978).
18) Minsky, M.: 知識を表現するための枠組，コンピュータビジョンの心理， pp. 238-332 (P. H. Winston 編，白井，杉原訳)

(昭和56年5月15日受付)
(昭和56年9月7日採録)