第 5 章

物体間の位置関係に関する空間推論の導入の提案

本章では，複雑なオクルージョンを含む室内シーン画像に対する画像認識システムを提案する．従来のシステムでは，物体が十分に画像中に表れていないと認識ができず，室内シーンのような複雑なオクルージョンを含む画像に対して対処できなかった．それに対して，我々の提案するシステムでは，物体が物体の上に載っているという関係である物体間の支持関係を定性的に推論することによって，他の物体によって隠されている物体の認識を可能としている．具体的には，最初に画像中に明確に表れている対象に対して３次元構造モデルを当てはめることによって物体の３次元構造を推定し，次に推定された物体の３次元構造を利用して，物体間の支持関係をチェックすることによって，部分的にしか見えていない物体の存在を推定したり，実在しない物体の候補を消去し，最終的に全体として整合性のとれた認識結果を得る．我々は，こうした認識を第4章で提案したマルチエージェント型の画像認識システムを拡張することによって実現した．本章では，システムについての詳細と，実際にインプリメントしたプロトタイプシステムによる実験，結果について述べる．

5.1 はじめに

通常，室内シーンにおいては，床の上に机があって，その上に計算機があるというように，複数の物体が積み重なって存在している (図 5.1)．そのため，手前にある物体が奥にある物体を隠してしまうというオクルージョンが発生しており，室内シーンを認識する場合はオクルージョンに対応することが必要である．従来のオクルージョンを含むシーンに対する認識は，予め正確な形状モデルが既知である物体を対象に行われることが多く，部分的な特徴にモデルを当てはめて，隠れている部分を推定することを行っていた．こうした方法は，工業部品などの予め形状が決まっている物体については有効であるが，一般のシーンを対象にする場合，物体の多様性の問題，つまり，同一の一般名称を持つ物体であっても形状は様々で，物体の正確な形状モデルが既知であることは通常はあり得ないために，有効でない場合がほとんどである．そこで一般のシーンに対しては，対象物体の典型的な構造特徴を記述するモデルを用いることになるが，図 5.1のようなオクルージョンが

図 5.1 室内シーン画像の一例.

多く発生している一般の複雑なシーンの場合には，オクルージョンのない物体を除いては物体の一部分を検出することさえ難しい.

本章では，従来のオクルージョンを含むシーンに対する認識においては物体単体を認識の対象としていたためにあまり利用されていなかった物体同士の位置関係を効率的に利用することによって，正確な物体形状が与えられていない状況での複雑なオクルージョンを含んだシーンに対する認識方法を提案する.

第4章で実現したシステムや，過去の一般の実世界シーンを対象とした画像理解システムの一部，例えば，The Schema System[31] などは，シーン中に含まれる物体同士の位置関係の情報を認識の手がかりとして用いていた．しかし，対象が風景画像であったために室内画像ほどオクルージョンは多くなく，またオクルージョンがあっても，認識の対象が「道路」「空」「木々」などの物体の構造や形状よりも色やテクスチャを手がかりとして認識する対象であったので，オクルージョンが問題になることはなかった．また，認識対象が色やテクスチャを主な手がかりとして認識する対象であったため，対象とするシーンは３次元シーンではあるが，基本的には領域分割と各領域に対するラベル付けという形で認識が行われており，認識は２次元的に行われていた．このため，物体間の位置関係の判定は単純に画像上での上下左右で行われていたが，それである程度うまく行っていた．こうした３次元シーンに対する２次元的な認識は，屋外の遠景画像のように広範囲のシーンを写した画像の場合は，認識対象となる物体自体の奥行きがシーンの奥行きに対して比較的小さいために有効である．ところが，室内画像の様に近景の画像の場合は認識対象となる物体の奥行きが無視できないので，２次元的な認識では不十分であり，３次元的な位置関係を推定することが不可欠である.

それに対して，人間は実世界に存在する物体の大まかな３次元構造を知識として持っているために，単一の画像からでも物体間の３次元的な位置関係をある程度推測することができる．また，物

体の構造に関する知識に加えて，物体は支えがないと下に落ちるといった物理法則の定性的な知識を持っているので，例えば，机の足が見えていなくても，平面があってその上に計算機が見えれば，机が計算機を支えているということが推測できる．そして，さらに机には適当な長さの足があって，足の下には床があって，机を支えているということも，画像中からボトムアップ的に認識することは困難であっても，知識から推測することが可能である．2次元の画像から3次元世界の構造を認識するシステムが，こうした実世界の定性的な物理法則に基づく3次元推論の能力を持つことは，より自然な画像の解釈を実現する上で必要な能力であると考えられる．

そこで，本研究では，従来の領域分割とラベリングによる2次元的な認識ではなく，物体の機能を反映した構造モデルの定性的モデル当てはめによる物体の定性的3次元構造の推定と，その結果に基づく物体間の支持関係に関する推論を行うことによって，オクルージョンを多く含むような室内画像に対する認識システムの提案を行う．最初のモデルの当はめは，画像から抽出したエッジ，領域をグループ化した画像特徴に対して行う．本研究では，同一種類であっても多様な形状，多様な見え方を持つ実世界の物体により広く対処できることに重点を置いているので，こうした同一物体の多様な見え方に対応できる方法を採用している．そして，物体が他の物体の上に載っているという関係を表す物体間の支持関係の推論によって，実在しない物体の候補を消去したり，部分的にしか見えていない物体の存在を推定したりすることが可能となり，モデル当てはめによる物体構造の推定の不正確さを補うことができる．この様にして，正確な物体形状が与えられていない場合でも，複雑なオクルージョンが発生している室内シーンの単一画像を認識することが可能となる．

関連研究としては，ある物体が他のある物体によって支えられているといったような物体間の力学的物理法則を考慮してシーンを理解する Cooper らの研究 [110] や Brand による研究 [111] がある．これらの研究では，定性的な力学法則の知識に基づいて画像によるシーンを解析を行っている．ただし，これらの研究では物体の認識が目的ではなく，物体の物理的作用の画像からの理解に焦点が当てられている．そのため，対象となる物体は，ブロックなどの単純な物が多く，本研究のような実画像を対象としたものではない．

本章では，まず，定性的モデル当てはめによる物体個々の認識と，その結果に基づく物体間の支持関係に関する推論について述べ，続いて，物体間に通常考えられる関係を予め記述した関係知識と，それを利用した物体の候補の評価値の計算について述べる．そして，次に，第4章において提案したマルチエージェントによる画像認識システム構成法 MORE (multi-agent architecture for Object REcognition)[3, 2] によるシステムの実現について述べ，最後にプロトタイプシステムによる動作例と 20 枚の画像に対する実験結果について述べる．

5.2 物体個々の認識方法

「机」「椅子」などの一般名詞で表現される物体は，同一種類であっても様々な形状を持つために，正確な形状3次元モデルを予め用意して置くことは不可能である．そこで，本研究では，モデルは人工物であれば物体の機能など認識対象の本質を表しているような構造 [37, 73]，例えば，椅子なら座面と足，机なら机上面と足などを表現する様にし，同一種類の物体でなるべく共通となる

ようなプロトタイプモデルを用意する．そして，画像から得られる物体の部分的な特徴や支持関係から予想される特徴に対してモデルを当てはめることによって，物体の存在を予想すると同時にその物体の定性的な 3 次元構造を推定する．モデルは物体によっては複数個用意して，その場合はその中で後述する画像特徴評価値がもっとも高いものを選択することにする．こうしたモデルの当てはめによる認識では，広い範囲の物体を認識することが可能になる代りに，異なる種類の物体間での区別が難しくなり，物体間で認識結果の競合が起こる場合がある．そうした場合は，モデル当てはめの正確さに加えて，後述する他の物体との関係も加味した上で競合を解決して，対象の同定を実現する．

なお，第 4.9節でも，物体個々を認識する認識モジュールの構築方法を述べたが，本章でも基本的には同様の方法で，人手による構築である．第 4.9節と異なるのは，定性的支持関係の推論と後述する新しい評価計算方法のためにモデルの定義が明確化されたことである．

5.2.1 モデルの表現

モデルは，多角形，線分で表現される**モデル要素** (model element)，及びモデル要素同士の接続関係を表現する**モデルグラフ** (model graph) によって定義される (図 5.2(a)(b))．図 5.2の机の例では，モデル要素は，4 つの頂点 (fl,fr,rr,rl) を持ち底辺と斜辺の長さがそれぞれ a, b である平行四辺形 (PG) と，2 つの端点 (t,b) を上下に持つ長さ c の垂直な直線 (VL) で，それぞれ，実世界中では，水平な長方形の面，垂直な棒であると定義されている．そして，平行四辺形の 4 つの頂点それぞれに垂直な直線の上の端点 (t) が接続していることをモデルグラフが表現している．この様に,「机」なら水平の机上面と垂直な足 (図 5.2(e)),「椅子」なら水平な座面と垂直な足というように，多くの種類の「机」「椅子」を代表するような典型的な構造をモデルとする．

図 5.2(c) では，後述する支持関係の推論のために，物体が他の物体の上に載っている時にその接面となると推定される支持必要面と，その物体が他の物体を上にのせて支持することができると推定される領域である支持可能面の情報を記述していて，システムがその物体の**支持必要要素** (to-be-supported elements) と**支持可能要素** (supportable elements) を推定できるようになっている．この例では，平行四辺形 (PG) の面すべてが支持可能要素，4 つの垂直な直線 (VL) の下の端点 (b) が支持必要要素であることを表している (図 5.2(f))．この 2 種類の要素の情報は，次節で述べる支持関係のチェックで用いられる．また，図 5.2(c) の 2 つの不等式は，モデル当てはめ時に使われる各モデル要素の画像中での大きさの比の範囲を表している．図 5.2(d) は，後述する画像特徴評価値の計算に用いられる各モデル要素についての重み値を定義している．

5.2.2 モデルの当てはめ方法

初めにシステムは与えられた画像から，モデル当てはめに必要な画像特徴として，エッジ，領域を抽出する．これらの領域，エッジはモデル当てはめの根拠になる画像要素なので，**根拠要素** (basis element) と呼ぶことにする．基本的には，根拠要素の抽出は既存のアルゴリズムを用いることとし，Canny edge detector[112] によるエッジ検出，Hough 変換による直線抽出，領域成長法による領域分割などのアルゴリズムを組み合わせて行う．そして，さらに perceptual grouping[64] の手法を用

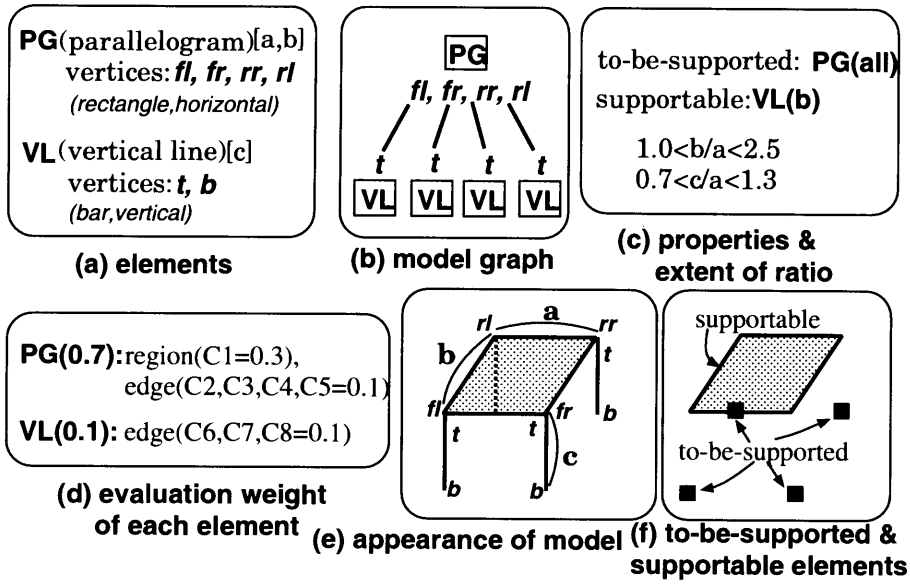

図 5.2 机のプロトタイプモデルの一例.

いて，抽出した直線をグループ化して，端点が近接している直線対，平行直線対，平行直線対に 1 本の直線が加わった U-shape，平行四辺形を抽出する.

次に，根拠要素に対して各モデルのモデル要素を定性的に当てはめることによって，物体の存在を予想し，さらに単一画像に対する物体の 3 次元構造の推定を行う (図 5.3). 当てはめは，全部のモデルについて，モデル要素の最も大きい平面要素から行い，順次，小さい要素を当てはめていく. 当てはめが成功するかどうかは，モデルに予め与えられている各要素の画像中での大きさの比の範囲を満たした上で，後述する**画像特徴評価値**がある一定の値以上の値になるかどうかで判断する. もし，当はめが成功すれば，**物体候補**が生成できたと見なす.

また，モデル要素の水平，垂直の属性，支持必要面，支持可能面の属性を用いて，物体のどの部分が水平，垂直で，どの要素が支持可能要素，支持必要要素であるかを推定できる. あくまでも定性的当てはめなので，定量的な正確さはないが，定性的推論に必要な程度の物体のおおよその大きさや向き，位置など推測することはできる. 以上のモデル当てはめに基づいて推定された物体の存在が予想される領域及びエッジをまとめて**候補要素** (candidate element) と呼ぶことにする. ここでの候補要素はオクルージョンが無い場合に本来見える物体全体の見え方を推定した場合の領域である.

モデル間での競合を解消する場合に用いられる評価値である**画像特徴評価値** V_{im} は 0 から 1 の間の値をとる値で，候補要素の各部分と根拠要素との対応の割合に応じて計算される. 根拠要素が候補要素に近いほど評価値が 1 に近くなるように，V_{im} を以下のように定義する.

$$V_{im} = min(\sum_{i=1}^{n} W_i \frac{b_i}{e_i}, 1) \tag{5.1}$$

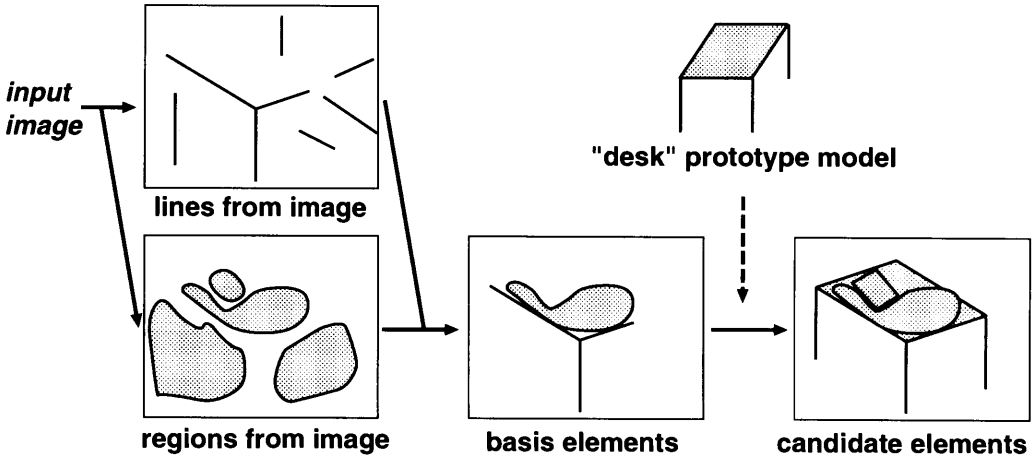

図 5.3 抽出された特徴に対して，3次元構造モデルを当てはめて候補要素を推定する．

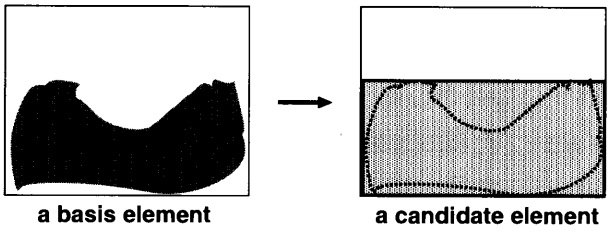

図 5.4 得られた根拠要素から,「床」の候補要素を推定する．

式中の n は合計のモデル要素数である．ただし，机上面のような平面で表現されるモデル要素は，内部の領域と輪郭線を別々に考えるので，図 5.2の机のモデルでは，机上面の領域，輪郭，4本の足の合計で $n = 6$ となる．W_i は各モデル要素の重要度を表現する重みで，本章における実装では机の場合は机上面の領域の重みを 0.3，輪郭を 0.3，足をそれぞれ 0.1 としている．b_i は各要素の根拠要素の画素数，e_i はモデル当てはめによって得られた対応する候補要素の画素数をそれぞれ表している．

なお，机，椅子などの物体とは別に，特定の形状を持たない平面的な「床」「壁」などの通常，背景となる物体は，図 5.4の様に得られた根拠要素に対して，平面的なモデルを当てはめることにする．これら背景物体の場合は，モデル要素は平面領域のみとなり，e を候補要素の画素数，b を根拠要素の画素数とすると，画像特徴評価値は

$$V_{im} = b/e \tag{5.2}$$

となる．

なお，前章の第 4.6.1節では，本章で述べた画像特徴評価値に相当する評価値を 5 段階の評価値の形状評価値としていたが，本章のシステムでは，評価値の計算方法を数式によって異なる定義をしたので，違う名前とした．

5.3 支持関係のチェック

支持関係とは，下にある物体が上にある物体を支えている関係のことである．実世界のすべての物体には重力がかかっているので，その下にその物体を支える他の物体がなければならない．人間はこうした物理法則を経験的に知っていて，それは無意識のうちに知覚に影響を与えていると考えられる．そこで，本システムでは，この物体が物体を支えるという関係を物体の認識に利用することとする．具体的には,「床」「壁」などの背景物体を除くすべての物体について，物体候補が生成されたら，その物体を支持することの出来る物体の候補が既に生成されているかどうか調べ，もしなければ支持可能な物体候補の存在を認識するように要求を出す．そして，もし最終的にその物体を支持する物体候補が見付からなければ，どの物体からも支持されていない物体は実在しないと見なして，候補を消去する．

支持関係のチェックは，支持必要要素の領域と他の物体候補の支持可能要素の領域が重複しているかどうかを調べることによって行う (図 5.5)．もし，支持可能要素領域が支持必要要素領域のほとんどを含んでいれば，支持可能要素を持つ物体が支持必要要素を持つ物体の下にあって支えているとみなし，その両方の物体の間には支持関係があるとする．支持関係の成立は，後述する関係知識とは無関係であり，2 種類の領域の関係のみで判断する．なお，図 5.5では，床が本を直接を支持しているとも判断できるが，物体候補 A が物体候補 B を支持していることを $A \Rightarrow B$ と書くとすると，床 ⇒ 机，机 ⇒ 本 という関係があるので，床 ⇒ 机 ⇒ 本 と判断できる．

もし，物体候補の支持必要要素に対して支持する物体候補がなければ，その支持必要要素を**仮想根拠要素** (virtual basis elements) として，支持する可能性がある物体の根拠要素の一部とみなすようにする (図 5.6)．そして，その仮想根拠要素を含む物体候補を生成可能であるかどうか，後述する関係知識にその候補を支持する可能性があると記述されている各モデルについて調べる．

図 5.6では，初めにワークステーション (以下，WS と略す) の候補が生成されて，その支持物体が存在しないので，WS 候補の支持必要要素を WS 候補を支持する候補の仮想根拠要素とみなす．そして，仮想根拠要素を根拠要素であると仮定して，仮想根拠要素の周辺部にさらに根拠要素となる画像特徴が存在するモデルを探す．この場合，机が WS を支持する可能性があるという関係知識が存在し，さらに，仮想根拠領域の周辺に机の根拠要素を見つけることができたので，両者を合わせて机の根拠要素として，モデルを当てはめることによって，机を検出することが出来た．

このように仮想根拠要素の考え方を導入することによって，上に物体が載って大きなオクルージョンが発生しているために認識不可能であった物体が認識可能となる．

5.4 関係知識 と 物体候補の評価

5.4.1 関係知識

第 4.6.2節で述べたようにシステムは，自分の物体と他の物体の間の通常考えられる関係についての知識，関係知識を予め持っている．これは必ずしも成り立つ必要はないが，多くの場合成り立っている関係で，物体候補の関係に関する評価，及び支持物体の探索に利用される．関係知識に記述

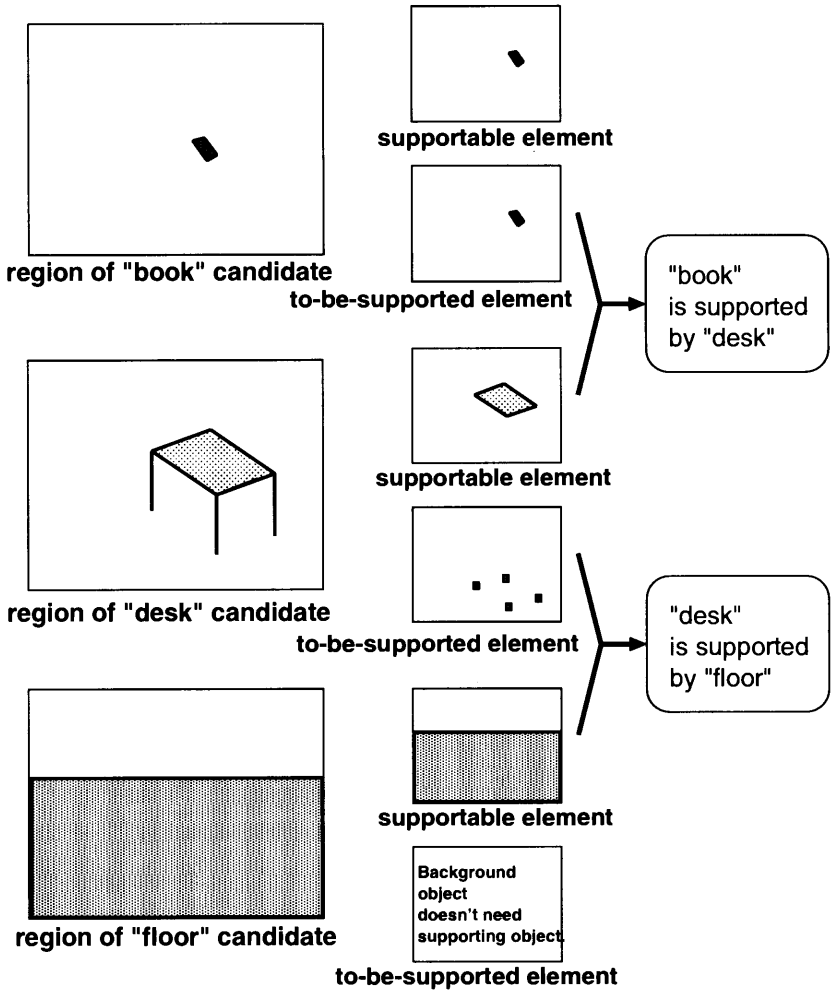

図 5.5 「支持関係」のチェック.

されている関係が周辺の物体候補との間に成り立っていると，その物体候補の関係についての評価値が高くなる．本システムで用いられている関係知識は 2 物体間の相対的な関係を記したもので，予めシステムに与えておく．具体的には 2 物体間の位置関係である．これらの関係はすべて定性的なものとして表現される．

関係知識は「関係（物体名，物体名）」という形で表現される．例えば,「本は机の上にある」「椅子と机は同じ平面上にある」という関係は，“on(book,desk)”, “next-to(chair,desk)” というように表現される．本章で実現したシステムでは，この 2 種類のみを利用している．

関係知識の評価は，物体候補と他の物体候補の間に，関係が成立しているかどうかチェックすることによって行なう．“on(A,B)” の場合，物体候補 A を支持している物体候補 B があれば，その関係知識が成り立っているとみなす．“next-to(A,B)” の場合は，物体候補 A と物体候補 B が同じ物体候補に支持されていれば，つまり，同じ支持平面上に載っていれば，関係知識が成り立っていると見なす．

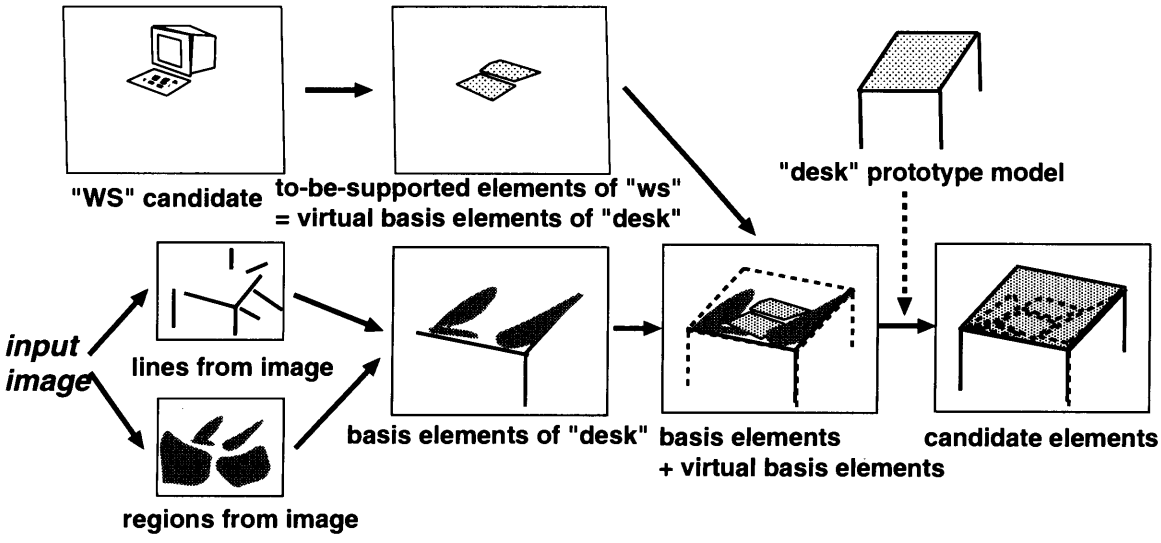

図 5.6 仮想根拠要素と通常の根拠要素を統合してから，モデル当てはめを行うことによって支持物体を認識.

ある一つの物体候補について，その物体と他の物体の間に成立している関係をすべて調べて，各関係の成立した数に基づいて計算したものが，**関係評価値** V_{re}(式 5.3) となる．関係評価値の計算は，画像特徴評価値と同様に第 4.6節で述べた方法を変更して，V_{re} が 0 から 1 の間の値をとるようにした．関係評価値は，物体候補が他の物体候補との間にどの程度通常成り立っていると期待される関係が成立しているかを示した評価値であり，その物体候補のシーン中での存在の自然さを表現している値であるといえる．関係評価値の計算式 (式 5.3) は，成立した関係の数が 0 個と 1 個では関係評価が大きく違うが，5 個と 6 個だとそれほど大きな違いがないということを反映した式となっている．

$$V_{re} = 1 - exp(-k \sum_{i=1}^{r} C_i n_i) \tag{5.3}$$

r は関係の種類数，C_i は関係 i についての予め決められている重みで関係の重要度を表現している．on の場合 1.0, `next-to` の場合 0.5 に設定している．n_i は関係 i について成立した数をそれぞれ表す．k は定数であり，現在の実装では実験から求めた値である 0.4 に設定している．

5.4.2 物体候補の評価

第 4章での方法では，形状評価値，関係評価値，面積を順番に比較するという単純な方法であったが，本章では計算式によってそれらを統合した値である**候補評価値**を求め，その値によって比較することとした．候補評価値 V は，画像特徴評価値 V_{im} と関係評価値 V_{re}，候補要素の総画素数 S から次式によって計算される．

$$V = (V_{im} \times S' + V_{re} \times w)/(S' + w) \tag{5.4}$$

$$S' = \min(S, 2w) \tag{5.5}$$

w は，画像特徴評価値と関係評価値の重みのバランスを決める定数で，候補要素の画素数が $2w$ 以上の時は V_{im} と V_{re} の重みが $2:1$，それ以下の場合は $S:w$ になる．w は現在の実装では 2500 に設定している．

候補評価値 V は，複数の物体候補の仮想根拠要素を除く根拠要素が重複して，競合が起こった場合の競合解消に用いられ，競合物体の中で候補評価値 V が最も大きい候補を最終的な候補として採用する．一方，評価値が小さかった候補は取り消しになる．

なお，評価値の計算においては，いくつかの定数パラメータを用いている．これらのパラメータを変更すると，評価が同程度の候補の競合解消の結果は変ってしまう可能性はある．しかし，その場合は画像上で見ても紛らわしい場合であり，明らかに一方が誤っている場合については，評価値が大きく異なっているので，パラメータの多少の変化によるは影響は少ない．

画像認識システムにおける物体候補の評価値の計算方法については，確立された方法の存在がなく，様々なシステムで様々な方法が用いられている [26]．ここでは，我々は D.Kim らの研究 [113] で用いられている計算式を改良し，我々のシステムに適用した．

5.5 システムの概要

5.5.1 システムの基本構成

システムは，第 4章で我々が提案したマルチエージェントによるシステム構成法 MORE (multi-agent architecture for Object REcognition)[3, 2] に基づいて構築する．システム構成法 MORE では，システムは単一種類の物体のみを認識する複数のエージェントのみから構成され，中央管理機構は存在しない (図 5.7)．そのため，エージェントの追加によりシステムを拡張することが可能であり，また，エージェント毎に異なる知識表現，認識手法を用いることができるために，大規模な画像理解システムの構築に向いている．それに加えて，処理の流れが固定されていないために，トップダウン処理とボトムアップ処理を柔軟に融合できるという特徴もある．

各エージェントは，単一クラスの物体を認識する認識モジュールと，エージェント間での協調を行う通信モジュールの 2 つから構成される．

認識モジュール (recognition module) は，物体モデルを持っており，画像中に含まれるある単一種類の物体を認識する．認識モジュールは，新しく候補を発見する度に，根拠要素，候補要素，支持必要要素，支持可能要素，画像特徴評価値などの情報を通信モジュールに送信する．

通信モジュール (communication module) は，認識モジュールの認識結果がシステム全体で整合性が保たれるように他のエージェントに結果の提示と競合解消のための交渉を行なう．通信モジュールは，認識モジュールが生成した物体候補の情報を受け取って，システム内で矛盾した結果が存在しないように，互いに情報を交換し合い，常にシステム内のすべてのエージェントの認識結果の整合性が保たれるように処理を行なう．また，通信モジュールは支持関係のチェックや，保持している関係知識を利用して関係評価値の計算を行う．

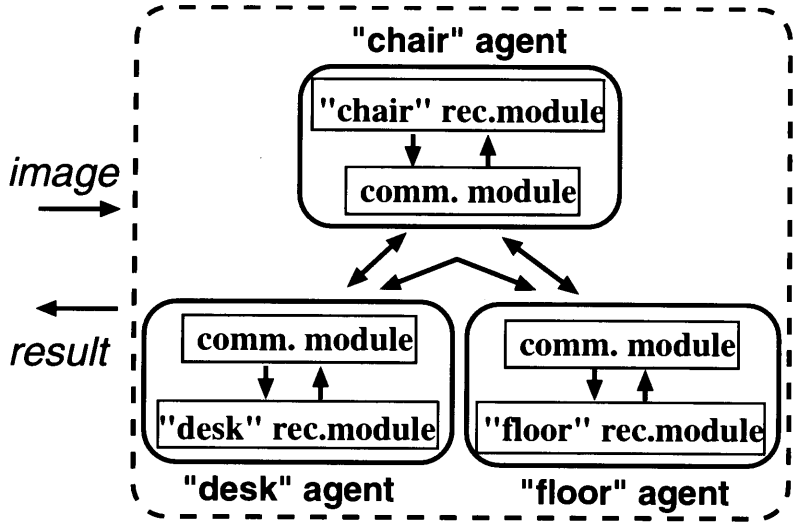

図 5.7 システムの基本構成.

5.5.2 システムの動作の概要

エージェントおよびそれを構成するモジュールの動作は，基本的には第4章で述べたものと同一であるが，本章で拡張した「支持関係」に関する部分は異なるので簡単に説明する.

初めに認識対象の画像が全エージェントの認識モジュールに送られ，通信モジュールが「初期認識要求」を認識モジュールに送る (図 5.8(1)). 認識モジュールは動作を開始し，1つ物体を認識する度にその認識結果を物体候補として通信モジュールに送る (同 (2)). それを受け取った認識モジュールは，他の全エージェントに対してその物体候補の情報を送信する (同 (3)). さらに，物体候補が背景物体でない場合に，他の物体の候補情報と照合した結果，支持物体を発見できないなら，仮想根拠要素を生成して，その情報を「支持要求メッセージ」として送信する (同 (4)). 物体候補の情報を受け取った他のエージェントの通信モジュールは，それが既に認識されている自分の物体候補の根拠要素と重複がないかチェックする. もし重複があれば，そのエージェントは「異義メッセージ」を返信し (同 (5)), 両エージェントの間で競合解消の処理が行なわれる. また,「支持要求メッセージ」と仮想根拠要素の情報を受け取った時は，その情報を認識モジュールに送って，再びモデル当てはめの処理を行う (同 (6)).

競合解消によって候補が消去されると，通信モジュール内にその競合候補の識別番号を記憶しておいて，競合候補がさらに別の候補に取り消された場合，もしくは，関係評価値が上昇して競合解消の結果が逆転した場合に，その候補を復活させることにする. こうして，常に互いに整合性のとれた認識結果のみを残すようになっている. この様に，常にシステム全体で認識結果の整合性がチェックされ，やがて，すべてのエージェントの認識が終了し，メッセージ待ち状態になると，システム全体の認識が終了する.

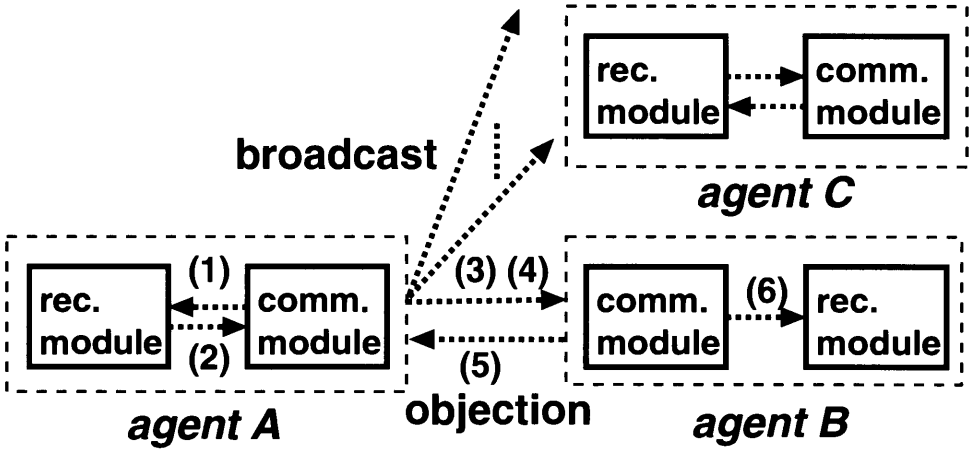

図 5.8 メッセージの流れ．(1) 初期認識要求．(2) 物体候補の情報．(3) 物体候補の情報のブロードキャスト．(4) 支持要求メッセージ．(5) 異義メッセージ．(6) 再認識要求．

5.6 実験

「机」「椅子」「床」「本」「ワークステーション (以下 WS と略す)」「壁」の 6 種類のエージェントを構築し，6 台の PC (Intel Celeron 450MHz) からなる PC クラスタ上に PVM [114] を用いて，プロトタイプシステムを実装した．プロトタイプシステムの実装においては，負荷分散については重点を置いていないので，1 エージェントを 1PC として実装し，各認識モジュールは独立に根拠要素の抽出の処理を行っている．根拠要素の抽出の処理を認識モジュール同士で共有することは可能ではあるが，現在の実装では行っていない．本節では，比較的単純な室内シーンとやや複雑なシーンの 2 枚の画像に対するシステムの動作の説明と，20 枚の室内画像に対する実験結果を示す．

5.6.1 動作例

システムは画像 (256 階調濃淡画像) が与えられると，画像を各エージェントの認識モジュールに画像を送信する．そして，認識モジュールがエッジや領域などの画像特徴の抽出を開始する．入力画像としてサンプル画像 1 (図 5.9, 480×360) が与えれたとすると，まず，各エージェントは，エッジ抽出，直線抽出，領域分割などの特徴抽出処理によって，直線エッジ (図 5.10) や分割領域 (図 5.11) が得られる．次に，直線エッジをグループ化して，端点が近接している直線対，平行直線対，平行直線対に両端点を結ぶ 1 本の直線が加わった U-shape，平行四辺形を抽出する (図 5.12)．すると，WS エージェントは，最も顕著な特徴として，WS のディスプレイの表示部分の平行四辺形を 2 つ抽出し，領域分割の結果からも同じ位置に平行四辺形領域を抽出する．そして，さらにその手前のキーボードが 1 つは平行四辺形として，もう 1 つは U-shape として抽出する．次に，これらの特徴に対して，平行四辺形のキーボード，左右の辺が画像中で垂直方向である平行四辺形のディスプレイの前面を当てはめることによって，WS 候補を生成する (図 5.13)．WS 候補の右側面はディスプレイ前面の右奥にある垂直直線エッジから推定することが出来た．そして，推定したディスプ

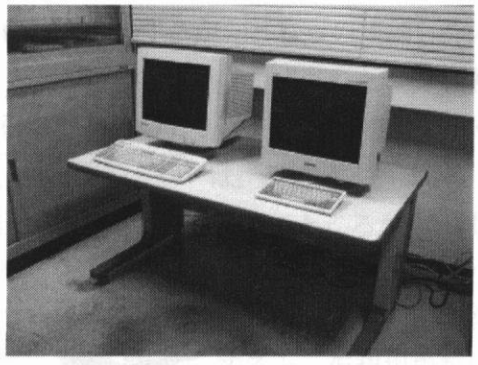

図 5.9 サンプル画像 1.

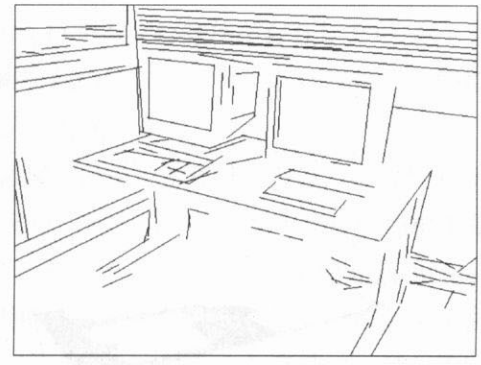

図 5.10 直線エッジ.

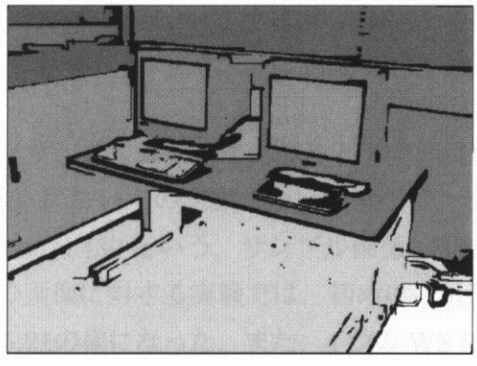

図 5.11 領域分割の結果.

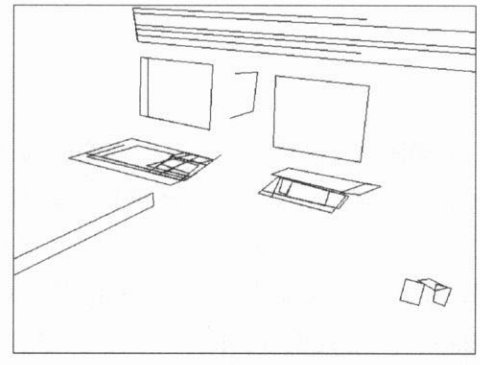

図 5.12 直線エッジからグループ化によって抽出した平行四辺形及び U-shape.

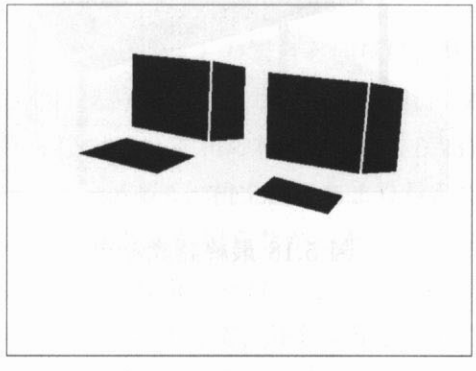

図 5.13 WS 候補.

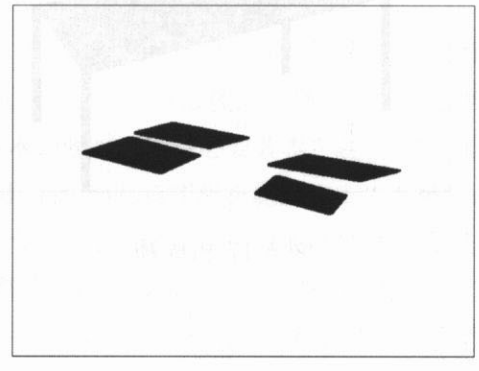

図 5.14 WS の支持必要要素.

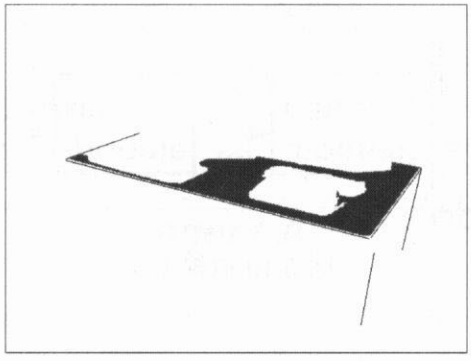

図 5.15 机の根拠要素.

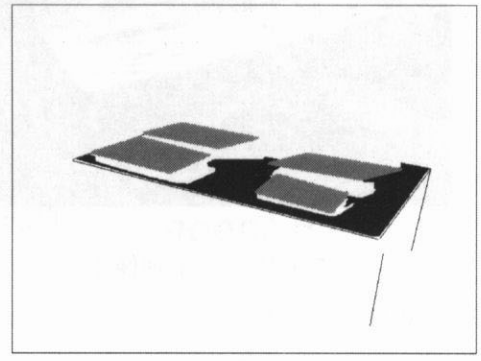

図 5.16 机の根拠要素と仮想根拠要素.

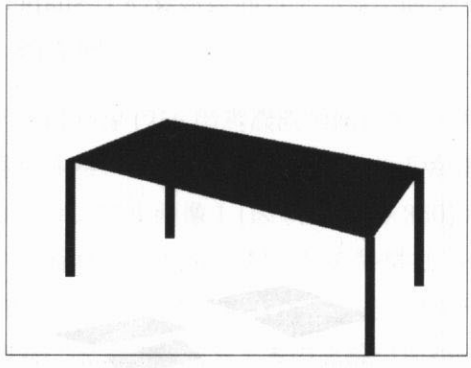

図 5.17 机候補.

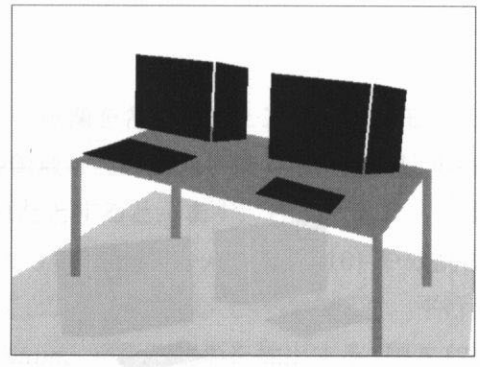

図 5.18 最終認識結果.

レイの前面と側面の四辺形の下部の辺から，その2辺をもつ平行四辺形領域を推定して，その領域とキーボードの領域を，支持必要要素の領域とする (図5.14)．こうして，WSエージェントは2つのWS候補を生成するが，どちらの候補もその候補を支持する支持物体が存在していない．そこで，WSエージェントは新規に生成されたWS候補の情報を他の全エージェントに送信する時に，同時に仮想根拠要素の情報を含んだ「支持要求メッセージ」も送信する．

一方，机エージェントは，最初は直線エッジと分割領域の画像特徴だけからは，特徴が十分でなく，机を検出することが出来ない．ところが，しばらくすると，WSエージェントから仮想根拠要素の情報を含んだ「支持要求メッセージ」が送信されてくる．机エージェントは，関係知識 `on(desk,WS)` を持っているので，WS候補の支持必要要素を仮想根拠要素とみなして，仮想根拠要素を含む机候補を認識しようとする．すると，仮想根拠領域の周辺に，図5.15のような直線エッジと領域を見付け，それを根拠要素として，仮想根拠領域と併合することによって，図5.16のように全体として十分な根拠要素を発見できる．そして，この根拠要素に対して，モデル当てはめを行うことによって，机候補を認識することができる (図5.17)．また，さらに，机候補からの「支持要求メッセージ」によって，床候補が正しく検出される．最終的には，図5.18の様に2つのWS，机，床が認識された．

次に，サンプル画像1に比べるとやや複雑なシーンの画像の例として，サンプル画像2 (図5.19, 640×480) についての認識について述べる．サンプル画像2は複雑なシーンの画像であるため，ここでは640 × 480という，サンプル画像1(480 × 360) に比べて解像度の高い画像を用いることとする．この画像に対する実験では，初めに，机上の4つのWS(図5.20) が認識され，その支持必要要素は図5.21の様になった．また，4つのWSの認識とほぼ同時に本の候補も認識され，候補が10個生成された (図5.22)．そのため，右から2台目のWSを除く3台のWSと本の間に競合が起きているが，競合解消の処理によって，3つともWSが残り，本が消去された．例えば，最も右のWSのキーボード部分と本との競合では，それぞれ画像特徴評価値 V_{im} が 0.78, 0.59，関係評価値 V_{re} が 0.33, 0.33 となり，候補評価値 V は 0.62, 0.43 となったために，本候補が消去された．なお，右から2つめのWS候補はキーボードの位置が誤って認識されたため，本来のキーボードが本として認識されてしまっている．

また，サンプル画像1の場合と同様に，WS候補，競合解消で残った本候補には支持物体が存在しないので,「支持要求メッセージ」が発行されて，机エージェントは仮想根拠要素であるWSの支持必要要素 (図5.21) を机の根拠要素 (図5.23) および本の候補の支持必要要素と統合して (図5.24)，オクルージョンが多く，机上面があまり見えていないにもかかわらず机を認識することができる (図5.25)．ただし，競合が起らなかった7つの本候補のうち，3つについては支持物体を発見できず，最終的には支持物体なしとなって，消去された．実際，これらの本は誤って検出されたもので，支持関係のチェックによって，正しく消去された．最終的には，図5.26に示す様に，後方右の机上に誤って2つの本が認識された以外は，WS，机，本，椅子，床などがほぼ認識ができた．なお，ここでは，WSエージェントの認識モジュールを暫定的にWSをディスプレイとキーボードの組で認識するように実装したため，後方左の机の上にある2つのWS本体の箱は認識されていない．

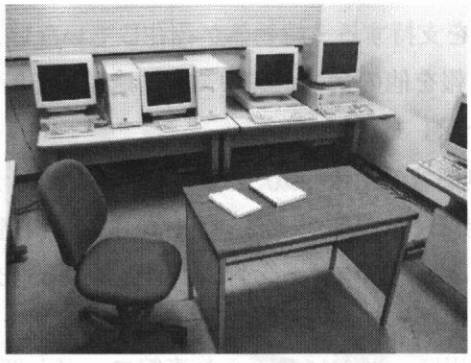

図 5.19 サンプル画像 2.

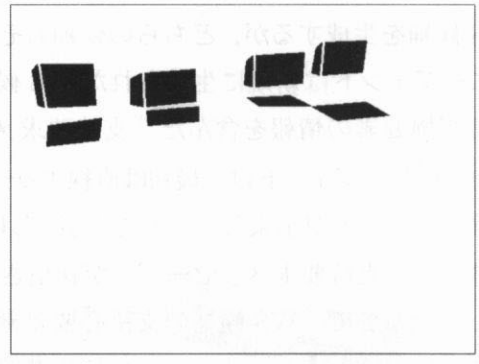

図 5.20 4 つの WS 候補.

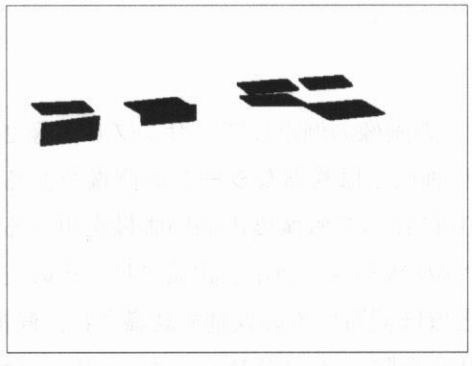

図 5.21 4 つの WS の支持必要要素.

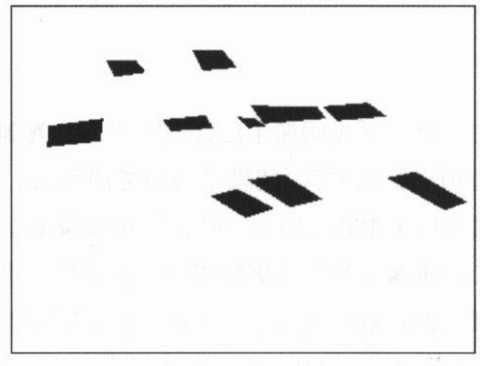

図 5.22 本候補.

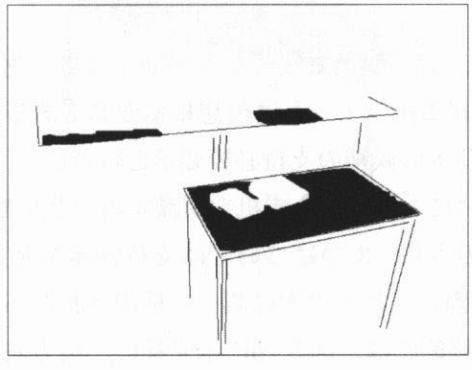

図 5.23 机の根拠要素.

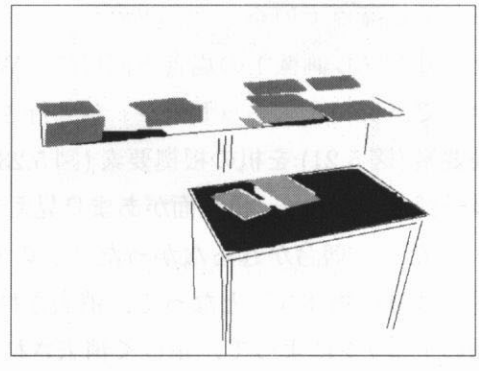

図 5.24 仮想根拠要素と根拠要素.

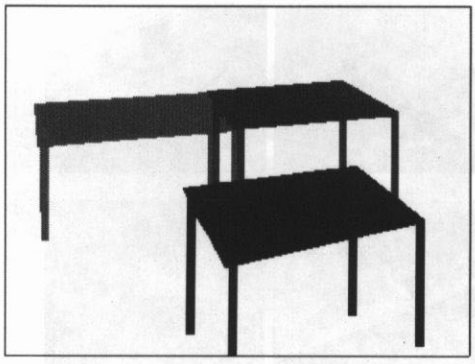

図 5.25 3つの机候補.

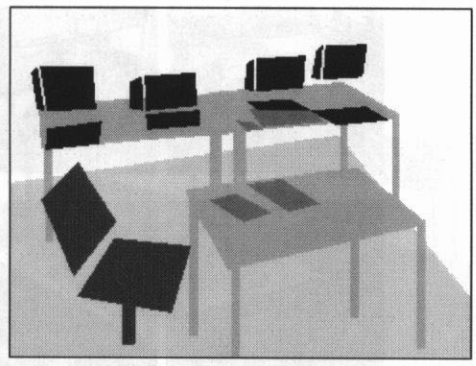

図 5.26 最終認識結果.

表 5.1 20 枚の画像 (480 × 360) に対する認識結果.

almost correct	half correct	almost incorrect
9	6	5

5.6.2 実験結果

20 枚の室内画像 (画像サイズはすべて 480×360) に対して，実験を行った．実験で用いた 20 枚の画像のうちの一部を図 5.27に示す．上段が単純な画像，下段は複雑な画像，中段は中程度の複雑さの画像を示しており，それぞれについて，7 枚 (サンプル画像 1 を含む)，7 枚，6 枚 (サンプル画像 2 を含む) 用意した．結果は「ほぼ認識出来ている (almost correct)」「半分程度認識出来ている (half correct)」「ほとんど認識出来ていない (almost incorrect)」の 3 段階に分けて評価した．それぞれ，画像中に含まれていて，対応するエージェントが存在する物体のうち，80%〜100%，30%〜80%，0%〜30%の物体が認識された場合に分類した．それぞれ，結果は 9 枚，6 枚，5 枚となった (表 5.1)．

5.6.3 実験結果に対する考察

20 枚の室内画像に対する実験において,「ほぼ認識出来ている」画像としては，先に説明したサンプル画像 1 (図 5.9) がその一例である．支持関係と仮想根拠要素を用いた認識によって，机上面が 2 台の WS によってほとんど隠されてしまっているシーンであるが机の認識が可能となっている．

次に,「半分程度認識できている」画像の例を図 5.28に示した．この画像では，机の上にあるノート PC，積み重ねられた本，本棚と並べられた本，広げられたノートなどが示されているが，認識結果では図 5.29に示したように，ノート PC が WS，広げられたノートが本と認識された．そのため，支持関係から，机の存在がほぼ正しく推定されている．しかし，本を平行四辺形として認識し

図 5.27 評価に用いた画像の一部．上段が単純な画像，下段は複雑な画像，中段は中程度の複雑の画像．

ているので，平行四辺形もしくは U-shape が検出されないと本を認識することができず，積み重なっている本や，本棚に入っている本は認識出来ていない．関係知識 `on(book,book)` を利用したとしても，一番上の本が認識出来ないと，その下の本も認識できない．この様な場合に対処するには，本が数冊まとまって存在している場合のモデルを用意するする必要があると思われる．

「ほとんど認識出来ていない」画像の例は図 5.30で，認識結果は図 5.31に示す．この画像は非常に複雑であり，最終結果では，床の一部しか認識できていない．WS のディスプレイは明確に写っているものの，机の上のキーボードは机の面と色が似ているために認識することが困難で，認識出来ていない．そして，そのため，WS の下の机も認識出来ていない．これは画像の解像度の限界に因るところが大きく，入力画像に解像度の高い画像，例えば 2400×1800 くらいの画像を利用して，初めは低解像度の画像を解析して，必要に応じて詳細な画像を利用するというような多重解像度解析を利用するのが望ましいと考えられる．また，中央の机の上には，モデルにない物体が置かれており，これも認識が不可能となっている．中央の机の最も手前に置かれているのは，定型がない包装材であるが，この種の物体は現在の認識方法ではモデルを用意することが難しい．動的輪郭やテクスチャ解析などの手法を取り入れることが必要であると考えられる．

5.7 まとめ

本章では，定性的モデル当てはめによって物体候補の定性的な 3 次元構造を推定し，さらに物体間の支持関係を確かめることにより，物体の候補の検証して，全体として整合のとれた認識を実現する方法について提案した．そして，さらに，こうした認識を実現するためのプロトタイプシステ

図 5.28 認識に部分的に成功した画像の例.

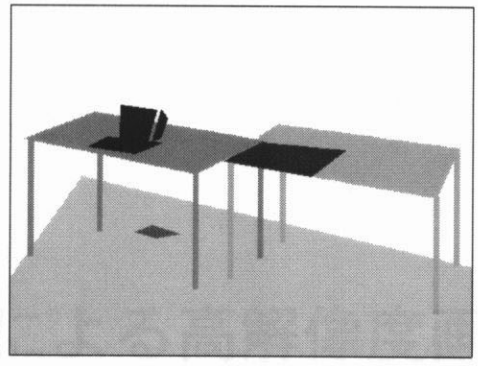

図 5.29 認識結果.

図 5.30 認識に失敗した画像の例.

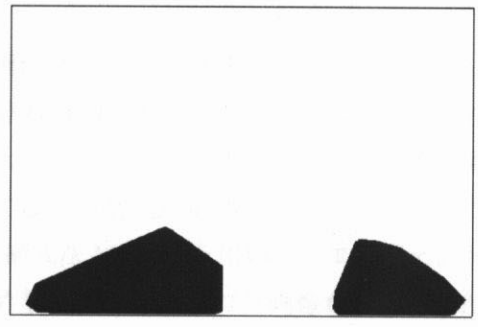

図 5.31 認識結果.

ムを我々が従来より研究しているマルチエージェント型の画像認識システムとしてプロトタイプシステムを実装し，実験により複雑なオクルージョンを含む室内画像に対応できることを示した.

現在のシステムでは，画像中の物体の大きさがある程度より大きくない場合，認識モジュールの3次元構造モデルの当てはめのうまくいかないことがあるので，今後の課題として，多重解像度解析を導入し高解像度の画像を入力画像とすることが挙げられる．なお，これについては，第6章でシステムを実現する．また，より実用的なシステムを目指すために，認識物体の種類を容易に増やすことが出来るように学習機構を取り入れていくことを検討する必要があるであろう.

第 6 章

多重解像度解析の導入による高解像度画像の利用の提案

本章では，100 万画素を越える高解像度の単一実画像に対して認識を行うシステムについて述べる．高解像度画像を認識に用いる場合，10〜30 万画素程度の画像を想定した従来のシステムにそのまま入力すると計算時間や必要記憶容量の著しい増大という問題が発生する．また一方，単純に縮小して入力することにすると有用な情報までを捨ててしまう恐れがある．そこで，本研究では，第4章で提案したマルチエージェント画像理解システム構成法 MORE を用いて，エージェントの協調作用により画像中の認識すべき部分を選び出し，予め数段階に縮小された画像から適切な解像度の画像を対象に応じて選択する機構を実現する．そして，その結果，処理時間をあまり増大させることなく，効率的に高解像度の画像を認識に利用することを実現した．本章では，そのシステムの詳細と，実験結果について述べる．

6.1 はじめに

近年，画像入力素子である CCD の性能の飛躍的な向上によって，1 辺が 1000 画素を越える全画素数 100 万以上の画像が容易に入手できるようになって来ている．特に最近では 400 万画素の CCD も一般的になり，デジタルカメラによって 2400 × 1800 の様な非常に解像度の高い画像を容易に手にいれることが可能である．けれども，そうした解像度の高い画像はデータ量が多く，そのまま画像認識処理を行うと計算時間の増大や記憶容量の不足などの問題が発生するため，従来の画像認識の研究では，縦横それぞれ数百画素程度で全画素数が数十万程度の画像を用いるのが普通であった．また，高解像度の画像を入力とする場合には，前処理の段階で画像を縮小することも多かった．画像を縮小するとデータ量は減るが，同時に高解像度の画像からしか得られない情報を捨ててしまっていることになる．そこで，初期段階では縮小した低い解像度の画像を利用してシーンのおおまか構造を認識して，部分毎に必要に応じてより高い解像度の画像を利用することによって，必要部分にのみズームをかけて，より正確に細部の構造まで認識を行う多重解像度解析を行うことが現実的である．そこで，本章では，単一の高解像度画像に対して，多重解像度解析を用いてシーン認識を

行うシステムを実現する．

必要に応じてではなく最初から高解像度の画像を利用すればよいという考えもあるが，エッジ抽出や領域分割などの特徴抽出アルゴリズムは場合によっては，n を画素数とすると計算量は $O(n^2)$ より大きくなることもあり，また，必要な記憶スペースも n に比例することになるので，通常の 16 倍の画素数がある 100 万画素以上もあるような高解像度画像をそのまま認識に用いるのは現実的でない．そこで，多重解像度解析を導入して，100 万画素以上もあるような高解像度画像を利用可能とすることによって，特徴抽出アルゴリズムの不完全さをカバーすることとする．つまり，画像全体の認識において重要な手がかりとなる大きな明瞭な候補を初めに検出して，それに基づいて，より解像度の高い画像に用いて，徐々に小さな候補を検出して行くような機構を実現する．

こうした複数の解像度の画像を利用して画像解析を行う多重解像度解析は 1980 年前後から主にエッジ抽出や領域分割などの画像処理の研究として始められ，その後画像認識システムにも応用されるようになった．Z.Li ら [115] や C.L.Tan ら [116] の研究では，高解像度の入力画像を数段階に縮小することによって，複数の解像度の画像を生成してピラミッド構造を作り，低解像度で大局的な構造を抽出し，それを基により高い解像度の画像に対しては必要な処理を行って，画像認識を実現していた．ただ，これらの研究では当時の計算機環境による制約のため，システムの実装は非常にプリミティブなものに留まっているという欠点があった．それに比べると現在では，計算機の性能が飛躍的に向上したので，より大規模で複雑なシステムを構築することが可能である．また，本研究と違って単一画像ではなくオンラインで画像を入力している場合には，アクティブビジョン [117] の枠組みにおいて，必要な部分にズームをかけることが研究されている．

本研究では，第 4章で提案した大規模なシステムの構築が比較的容易であるマルチエージェントによる画像理解システム構成法 MORE (Multi-agent architecture for Object REcognition)[2, 3] を利用して，実世界シーンの単一画像に対して，そのシーン中に含まれる物体の一般名称と，物体同士の位置関係を認識するための画像理解システムをエージェントの協調作用に基づく多重解像度解析システムとして実現する．従来の多重解像度解析を用いた物体認識の研究は主に単体の物体を認識する研究がほとんどであり，複数の認識対象が存在する実世界シーンの単一画像に対するマルチエージェント型のシステムで，多重解像度解析を実現したシステムは存在しなかった．

本章では，こうした高解像度画像に対する多重解像度解析をマルチエージェント型画像理解システムに導入することよって，記憶容量や計算時間などの制約で直接利用することのできなかった高解像度画像を効率的に認識に利用して，従来の低解像度画像のみからでは認識することのできなかった認識対象を認識できるシステムの提案を行う．そして，さらに，室内画像を対象としたプロトタイプシステムを複数の計算機からなる PC クラスタ上に実装することを試みる．

6.2 多重解像度解析

6.2.1 導入の構想

一般に画像認識システムでは，直線検出や領域抽出などの特徴抽出処理で，ある一定量の以上の特徴が検出されないと，物体の候補を生成するのは不可能である．そこで，初期段階の認識で認識

できなかった候補を他の候補との関係から推定して，トップダウン的に認識する「再認識」というメカニズムが存在している．これは通常，画像特徴抽出のアルゴリズムのパラメータを変更するか，または，画像特徴から候補を推定する処理において，候補と認める閾値を下げることで実現しているが，パラメータや閾値の変更の方法に明確な指針がないことが多く，また，画像中において不鮮明であったり小さかったりする候補はパラメータや閾値の変更によっても認識できないという問題点があった．そこで，本研究では「再認識」時により高い解像度の画像を必要な部分だけ利用することによって，初期認識時に認識できなかった候補を認識することとする．つまり，初期認識では縮小した低い解像度の画像を利用して，画像中で大きく明瞭に，またオクルージョンがなく手前に現れている物体をまず最初に認識して，次にそれらの初期認識で認識された物体候補を手がかりに，部分毎に必要に応じてより高い解像度の画像を利用することによって，未認識の物体を再認識することとする．

また，全体の認識の終了時に，ある一定の面積以上の未認識の領域が存在したら，その領域に対して，より高い解像度の画像を用いて「再認識」を行うメカニズムも導入する．

オンラインで画像を入力している場合は，アクティブビジョン [117] のように，必要な部分にズームをかけることが可能であるが，本研究では単一画像に対する認識を目的としているので，そのように必要に応じて動的に画像を取得することは不可能である．そこで，あらかじめ解像度の高い画像を用意しておいて，必要な部分のみ高解像度の画像を利用することにする．

こうした問題へ対処法としては，特徴抽出アルゴリズムの性能を高めるという方法と，より解像度の高い画像を利用して，拡大した対象を解析するという方法が考えられるが，本研究では後者のアプローチであるといえる．

6.2.2 画像ピラミッドの生成とレベル選択

初めに入力画像から多重解像度解析で必要な画像ピラミッドを生成する．入力画像がある一定画素数 (例えば，10 万) 以上なら，その画像の縦横それぞれの大きさを一定倍縮小 (0.5 から 0.75 程度) した画像を生成し，画像の大きさが一定値以下 (例えば，横 320 画素以下) になるまで繰り返す．そうすると，図 6.1に示した様な画像ピラミッドが得られる．ここでは，画素数が少ない順，つまり解像度の低い順に**レベル 0, 1, 2,...** の画像と呼ぶことにする．画像ピラミッドの生成は，システムの認識動作が始まる前に，各認識モジュールで行うこととし，認識要求時の画像レベル指定に従って認識に利用する画像レベルを切り替えることとする．

「再認識」時における画像解像度の選択は，高解像度画像を数段階に縮小することによって予め得られている画像群から，認識領域の画素数 (例えば 20 万画素) がある一定画素以下になる最も大きい画像を選択することとする．この閾値を変化させることによって，どの程度画像を詳細に時間をかけて認識するかどうかを決めることができる．以後，再認識時に選択される画像のことを「**適切レベル**」の画像と呼ぶことにする．適切レベルは次の基準によって選択する．

$$l_p = \arg\max_l \{S_i^l | S_i^l < t\} \qquad (0 \le l \le l_{\max}) \tag{6.1}$$

$l_{\max}$ は画像ピラミッドの最大レベル，S_i^l はレベル l の画像での領域 i の画像数，t は閾値をそれぞ

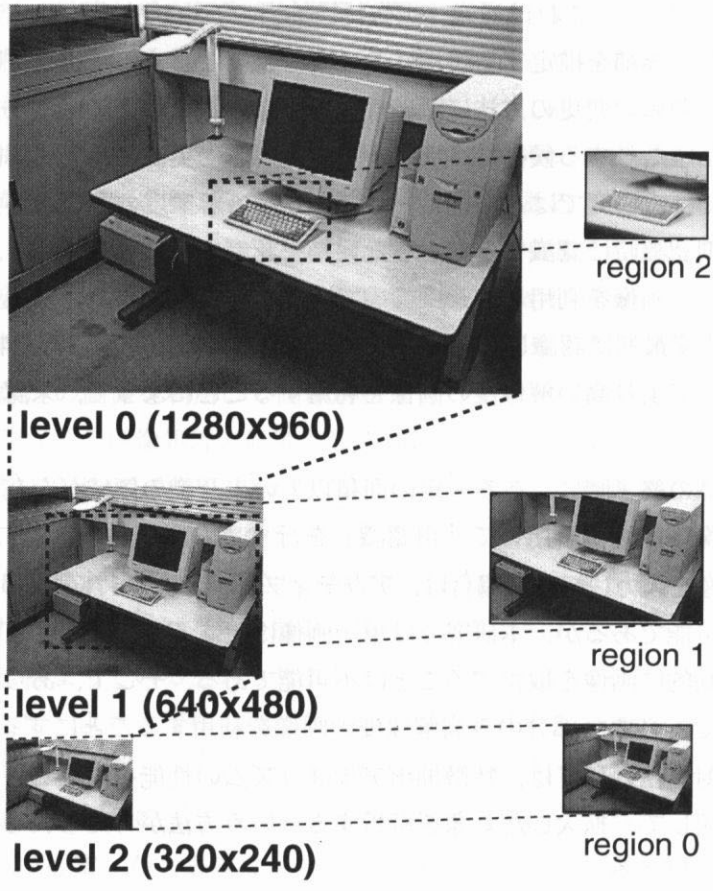

図 6.1 多重解像度解析のための画像ピラミッド.

れ表す.

例えば，入力画像の解像度が 1280×960 の場合，図 6.1のように，640×480，320×240 の画像を入力画像を縮小することによって画像ピラミッドを生成し，原画像から大きい順にそれぞれレベル 0, 1, 2 の画像と呼ぶことにすると，初めにシステムは最大レベルのレベル 2 の画像，つまり図 6.1中の region 0 を解析し，シーンの大まかな構造を得る．この場合は，机，床，机の上にあるディスプレイが認識される．次に，システムは床や机の上に他に物体が存在しないかどうか，一つ上のレベル 1 の画像を解析する．この場合は，レベル 0 で推定された床領域，机領域とその周辺領域の region 1 を解析する．この段階ではディスプレイの手前の机の上の領域に何か物体らしきものがあるということが分かるが，どんな物体であるか分からない．そこで，さらに高い解像度のレベル 0 の画像で，レベル 1 で物体の存在が推定された領域 region 2 に対して解析を行う．すると，その領域はキーボードであるということが分かる．表 6.1は，図 6.1における各 region の各レベルでの総画素数を示している．閾値 th を 200,000 にセットした場合，region 0, 1, 2 の適切レベルは，式 6.1 により，表中に下線で示した 2, 1, 0 がそれぞれ選ばれる．

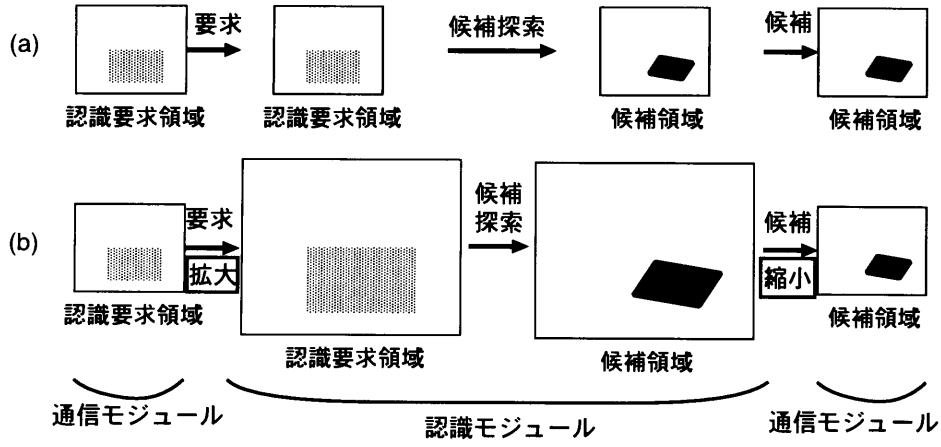

図 6.2 認識要求時の指定領域と候補領域の取り扱い．(a) レベル 0 の時．(b) レベル 1 以上の時．

6.2.3 実装上の工夫

実際に通信モジュールが扱う領域データはレベル 0 の大きさで表現することとして，レベル 1 以上の画像は各認識モジュール内でのみ，そのレベルの大きさの画像として扱われることとする．つまり，図 6.2に示すように，通信モジュールからの領域指定認識要求時の指定領域，認識モジュールから送られて来た候補は認識モジュール側ではレベル 0 の画像と同じ大きさで表現され，認識モジュール側では本来のレベル 1 の大きさで扱われる．この変換は，認識モジュールでの送受信時に行われ，そうすることによって，通信されるデータ量が増大するのを防ぐとともに，送信モジュール側ではレベル 0 の大きさの画像のみに対応したデータ構造を持っていればよく，従来システムからの大幅な変更は不要で，認識要求時に画像レベルを指定することを追加するのみでよくなる．

6.3 シーンの認識の方法

本章のシステムでは，図 6.1に示したような人工物で構成される実画像を対象としている．目的とするシーンの認識は，個々の物体とその物体同士の位置関係を認識することによって実現する．

表 6.1 各レベルにおける各領域の画素数．

region no.	level 0	level 1	level 2
0	1,228,800	307,200	**76,800**
1	616,224	**154,056**	38,514
2	**54,528**	13,632	3,408

6.3.1 個々の物体の認識

個々の物体の認識は，画像から得られる直線エッジおよび領域に定性的構造モデルを当てはめることによって行う (図 6.3)．基本的には，第 5.2節で述べた方法と同じである．

具体的には，最初に画像全体または一部分に対して，Canny edge detector や Laplacian zero crossing などの一般的なエッジ検出法を用いてエッジを検出し，次に Hough transform によって直線エッジを検出する．また，それとは別に，同じ画像に対して領域成長法などの一般的な領域分割法を適用して領域分割を行う．次に，得られた直線エッジと領域を利用して，定性的構造モデル (図 6.4) を画像に当てはめ，物体候補の存在領域を推定する．なお，構造モデル当てはめに用いられた画像特徴のうちエッジを根拠エッジ，領域を根拠領域とそれぞれ呼ぶこととする．図 6.3中には「含まれるべき領域 (region to be included)」というものが含まれているが，これは物体同士の位置関係による認識から，含まれるべき領域であると推定されたもので，再認識時のみ存在する．

定性的構造モデルは，平面と線分の組み合わせによって簡単に定義し (図 6.4(a)(b))，実世界において通常，それぞれの要素が水平か垂直かまたはどちらでもないかの属性を与える．そして，さらに物体同士の支持関係（物体が他の物体の上に載っているという関係）の推論のために，モデルには，その物体が他の物体の上に載っているときにその接面となると推定される支持必要面と，その物体が他の物体を上にのせて支持することができると推定される領域である支持可能面の情報 (c) を持たせ，認識モジュールがその物体の**支持必要領域**と，**支持可能領域**を推定できるようにする．これらの領域は画像中での領域である．

ここで用いる構造モデルは，人工物なら，物体の機能など認識対象の本質を表しているような構造，例えば，椅子なら座面と足，机なら机上面と足など [73] を表現する様にし，同一種類の物体でなるべく共通となるようなプロトタイプモデルを用意する．例えば,「机」は平行四辺形とその 4 つ頂点の下方に垂直な長さの等しい 4 本の線分を持つといったように，多くの種類の机を代表するような典型的な構造を表すようにする (図 6.3)．このプロトタイプモデルは物体によっては複数用意して，その場合は当てはめたときに最も後述する形状評価値が高くなるものを選択することにする．

モデルの当てはまり具合いを示す値である**画像特徴評価値** V_{im} は 0 から 1 の間の値をとり，画像レベルを考量して，候補領域の各部分と，根拠領域，根拠エッジとの対応の割合に応じて計算される．画像レベルが小さいほど，根拠領域，根拠エッジが推定領域に近いほど，評価値は 1 に近くなる (式 6.2)．

$$V_{im} = min((\sum_{i=1}^{n} W_i \frac{b_i}{e_i})^k, 1) \tag{6.2}$$

$$k = 1 + l/(l_{\max} + 1) \tag{6.3}$$

n は根拠領域，根拠エッジの合計要素数 (図 6.4の場合は $n = 8$) を表す．W_i は各要素の重要度を表す重み (図 6.4の場合は (d)) であり．これは予めモデル情報として与えられている．b_i は根拠領域，根拠エッジの画素数，e_i は推定領域，推定エッジの画素数，l は候補が生成された画像レベル，$l_{\max}$ は画像ピラミッドの最大レベルをそれぞれ表している．

例えば，机の場合は要素数は 6 で，W_i は机上面の領域に 0.3，輪郭に 0.3，4 本の足にそれぞれ

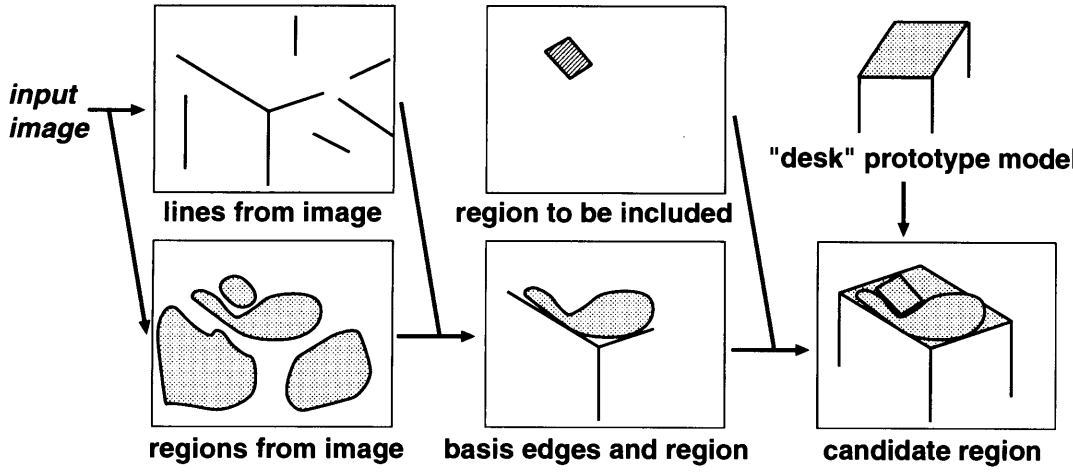

図 6.3 抽出された特徴から根拠エッジ，根拠領域を選び出し，それに対して 3 次元構造モデルを当てはめて候補領域を推定する．

0.1 を与えている．もし，候補がレベル 0 以外で生成された場合，V_{im} は重み付き合計値の k 乗となる．k は式 6.3 で計算され，$1 \leq k < 2$ となるので，重み付き合計値は 0 以上 1.0 未満の値をとり，候補が生成された画像レベルが大きければ大きいほど (つまり，解像度が低いほど)，V_{im} が小さくなる．この V_{im} は，後ほど説明する競合解消の処理に用いられる．

6.3.2 物体同士の関係の認識

第 5.2節で述べた方法と同様に，本章におけるシステムでも，すべての物体について，下にある物体が上にある物体を支えている関係である「**支持関係**」が成り立っているか調べることを行う．支持関係のチェックは，支持必要領域と支持可能領域が重複しているかどうかを調べることによって行い，もし，支持可能領域が支持必要領域のほとんどを含んでいれば，支持可能領域を持つ物体が支持必要領域を持つ物体の下にあって支えているとみなし，その両方の物体候補の間には支持関係があるとする．もし，1 つの物体候補に対して支持する候補が複数ある場合は，支持可能面の面積が小さい方と支持関係があるとみなす．

この支持関係のチェックによって，未認識の物体候補の存在を推定できたり，実際には存在不可能な物体候補を推定することができる．つまり，もし，どの物体にも支持されていない物体候補 (床や壁などの背景物体を除く) の下にはその物体を支持する物体があるはずであるし，最終的にどの物体にも支持されていない物体候補は基本的には誤認識であると考えることができる．どの物体にも支持されていない物体候補が認識された場合，後述する関係知識に基づいてその物体候補を支持する可能性のある物体について「再認識」**(支持物体認識)** が行われる．その場合，どの物体にも支持されていない物体候補の支持必要面が先に述べた「含まれるべき領域」となり，その支持必要面を含む支持可能面を持つ物体候補が探索される．

システムは個々の物体のモデルの知識以外に，物体と物体の間の通常考えられる位置関係につい

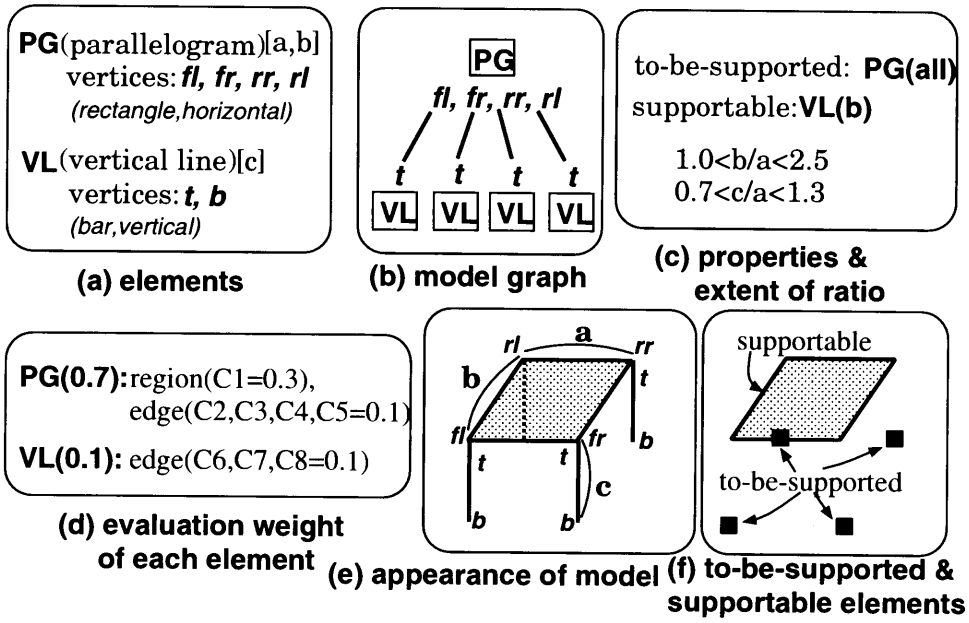

図 6.4 机のプロトタイプモデルの一例．(a) モデルの要素．(b) 要素間の接続関係．(c) 支持属性と要素長の許容範囲比．(d) 各要素の評価時の重み．(e) モデルの構成．

ての知識である**関係知識**を持っている．関係知識は「関係（物体名，物体名）」という形で表現される．例えば,「本は机の上にある」「椅子と机は同じ平面上にある」という関係は，“`on(book,desk)`”, “`next-to(chair,desk)`” というように表現される．本章のシステムでは，第5章と同様に，この 2 種類のみを利用している．

関係知識を用いることによって，先ほど述べた上に載っている物体候補から下にある物体を再認識を要求するメカニズムに加えて，逆の，下にある物体から上に載っている物体の再認識を要求するメカニズムも実現できる．つまり，物体候補が認識されたら，その物体の支持可能面とその周辺部分について，関係知識 “on” の成り立っている物体について「再認識」**(被支持物体認識)** を行うことで実現可能である．

関係知識の評価は，第 5.4.1節で述べた方法と同一である．簡単に述べると，物体候補同士の間で関係知識が成立するかどうか 1 つづつチェックすることで行い，すべての関係知識をチェックして，成立した関係それぞれについての重み値の合計が関係評価値 V_{rel} となる (式 6.4).

$$V_{re} = 1 - exp(-k \sum_{i=1}^{r} C_i n_i) \tag{6.4}$$

r は関係の種類数，C_i は関係 i についての予め決められている重みで関係の重要度を表現している．on の場合 1.0, それ以外の場合 0.5 に設定している．n_i は関係 i について成立した数をそれぞれ表す．k は定数であり，現在の実装では実験から求めた値である 0.4 に設定している．

6.3.3 競合の解消

同一の画像特徴から複数の候補が生成された場合，候補同士の競合が発生する．その場合，第5.4.2節と同様に，候補評価値 V を求めて，それを用いて，候補の比較を行い，V の小さい方が取消しとなる．**候補評価値** V は，画像特徴評価値 V_{im} と関係評価値 V_{re} の重み付き合計値として，次のように求める．

$$V = (V_{im} \times S' + V_{re} \times w)/(S' + w) \tag{6.5}$$

$$S' = \min(S, 2w) \tag{6.6}$$

S は候補要素の画素数を表す．w は，画像特徴評価値と関係評価値の重みのバランスを決める定数で，候補要素の画素数が $2w$ 以上の時は V_{im} と V_{re} の重みが $2:1$，それ以下の場合は $S:w$ になる．これは，生成された候補の画素数が少ない場合は，画素数が多い場合に比べて関係を画像特徴より重視するという考えに基づいている．w は現在の実装では 2500 に設定している．

なお，取消された候補は，競合した画像特徴がその候補の根拠エッジまたは根拠領域の一部分 (0.5以下) であれば，競合した特徴を含まない様に候補を変形する「再認識」**(候補変形認識)** を行う．この再認識では，競合した候補が競合要素を含まない様に修正する再認識を行う．

6.4 システムの基本構成

システムは，第4章で提案したマルチエージェントによるシステム構成法 MORE (Multi-agent architecture for Object REcognition) に基づいて構築する．MORE ではシステムは単一種類の物体のみを認識するエージェントの集合体として構築され，それぞれがある特定の種類のみの物体を画像中から認識するという目的を持つ．複数のエージェントが同じ画像要素をそれぞれ異なる物体であると認識した場合は競合が発生し，エージェント間で物体候補の評価値の比較を行ない競合を解消する．各エージェントは対象とする物体の個々の特徴に関する知識と，通常考えられる物体同士の位置や相対的な大きさなどの関係の知識を持っており，これらを統合して利用することにより，各エージェントの物体の認識及びエージェント間で解釈の矛盾が生じた場合の競合解決を実現している．

各エージェントは，定性モデル当てはめによる物体候補の認識を行う**認識モジュール** (recognition module)，物体候補オブジェクトの生成，他エージェントの生成した物体候補の監視，関係知識を利用した再認識の要求などを行う**通信モジュール** (communication module) の2つのモジュール以外に，本章では．物体候補が見つかると通信モジュールによって生成される**候補オブジェクト** (candidate object) を新たに導入した (図6.5)．初期状態では，認識モジュールと通信モジュールがそれぞれ1つずつ存在し，候補オブジェクトは存在しない．認識モジュールが物体の候補が1つ見つける度に候補オブジェクトを通信モジュールが1つずつ動的に生成される．これらのモジュールとオブジェクトはすべて並列に動作する．

さらに，第4章や第5章のシステムとは異なり，本章におけるシステムでは，画像処理モジュールをエージェント群とは別に独立させ，入力画像から直線エッジや領域などの画像特徴を抽出する

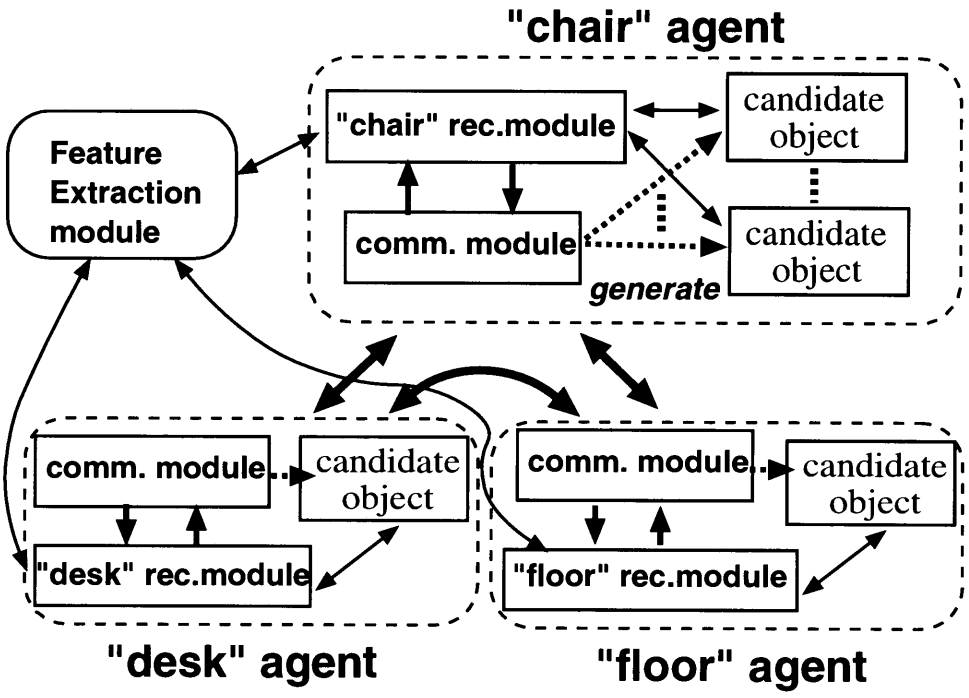

図 6.5 システムの構成.

画像特徴抽出モジュール (feature extraction module) を用意した．特徴抽出モジュールでは，入力画像から多重解像度解析で必要な画像ピラミッドを生成する．入力画像の縦横それぞれの大きさを一定倍縮小 (現在の実装では，0.7 倍) した画像を生成し，画像の大きさが一定値以下 (300 画素以下) になるまで繰り返す．そうすると，図 6.1に示した様な画像ピラミッドが得られる．画素数が少ない順，つまり解像度の高い順に**レベル 0, 1, 2,...** の画像と呼ぶ．

画像特徴抽出モジュールは認識モジュールや候補オブジェクトからの要求によって，指定された解像度の画像の指定された矩形領域から直線エッジ，平行直線対，多角形を構成する直線群，領域などを抽出する．このように，特徴抽出モジュールは認識モジュールの下請の役割を果たしているので，以後は認識モジュールに含まれていると見なして，詳しくは触れないこととする．

認識モジュールは，入力画像及びそれを縮小した 3 つのレベルの画像を持っており，通信モジュールや候補オブジェクトの要求に応じて，指定されたレベルの画像の指定された領域に含まれる単一種類の対象物体の認識を行う．エージェント内部の構成については特に規定はなく自由であり，それぞれの対象に適した認識手法，知識及びその表現方法を使用する．認識エージェントは，物体候補領域を画像中から抽出すると，その候補領域に対する確信度を自己評価値として付けて，領域情報と評価値を物体候補の情報として認識要求を発信した通信モジュールまたは候補オブジェクトに送信する．

通信モジュールは，初めに認識モジュールに初期認識要求を送って，その結果に基づいて，候補オブジェクトを生成する．その後は，他のエージェントの候補オブジェクトからの候補情報を自分

表 6.2 6 種類の認識要求.

認識要求名	領域指定	画像レベル	条件	探索物体
初期認識要求	画像全体	最大レベル	なし	すべて
候補変形要求	元の候補領域	適切レベル *1	指定領域を根拠要素として含まない	元の候補
候補更新要求	元の候補領域	適切レベル *2	なし	元の候補
支持物体認識要求	要求元候補の支持必要領域の周辺	適切レベル	要求元の支持必要領域を支持可能領域が含む	関係知識にある物体
被支持物体認識要求	要求元候補の支持可能領域とその上方	適切レベル	要求元の支持可能領域に支持必要領域が含まれる	関係知識にある物体
未認識領域認識要求	未認識領域	適切レベル *3	なし	すべて

*1. 元の候補が適切レベル-1 の画像で認識されたものなら適切レベル-1 で認識. *2. 元の候補が適切レベルで認識されたものなら適切レベル-1 で認識. *3. 同じ領域が既に初期認識か未認識領域認識で適切レベルで認識されていれば, 適切レベル-1 で認識. 既に適切レベル-1 で認識されていれば, 終了.

のエージェント内の候補オブジェクトに通知したり, 通信モジュール自身が持っている物体間の位置関係などの知識を記述した**関係知識**に基づいて, 認識モジュールに対して存在可能性の高い領域を指定して, 条件付き認識要求を行ったりする. また, 関係知識を用いて, 自分の物体候補と他の物体候補との関係に関する評価も行う.

新しく導入した**候補オブジェクト**は, 通信モジュールによって生成され, 一度生成されると自分の候補領域が他の候補との間で整合性が保たれるように他のエージェントに対して候補情報の提示と, 競合解消のための交渉を行う. また, 整合性を保つために, 認識モジュールに対して自候補領域の変形要求を行うこともある. 競合解消の交渉の結果, 候補自身が消滅する場合もあるが, その場合も候補オブジェクト自体は存在し続け, 常に復活の機会をうかがうことになる. 基本的には, 第 4 章, 第 5 章での通信モジュールが行っていた, 生成された候補情報の管理を代りに行っていることになる.

6.4.1 認識要求

通信モジュールが認識モジュールに認識要求 (recognition request) を送ると, 認識モジュールは認識処理をスタートさせる. 認識要求としては, 表 6.2に示す 6 種類が存在する. そのうち 1 つは「初期認識要求」, 残りは「再認識要求」である.

(1) 初期認識要求 (initial recognition request)

最初に発行される認識要求である. 最高レベル, つまり最も低い解像度の画像の全体に対して認識が行われる.

(2) 候補変形要求 (modification recognition request)

解消の後, 負けた候補を勝った候補との重複領域を領域を含まないような新たな候補に変形

させるための再認識を行う．負けた候補の領域に対して，適切レベルで相手候補の根拠要素を含まないような候補を探す．

(3) 候補更新要求 (renewal recognition request)
一度認識された候補で競合がない場合に，適切レベルで再認識する．

(4) 支持物体認識要求 (supporting recognition request)
もし，新しく生成された候補（背景物体を除く）が他のどの候補にも支持されていない場合に，支持物体を探す．適切レベルで，その候補の支持必要要素を含むような支持物体を再認識する．この認識要求は，関係知識に基づいて，支持なしの候補を支持する可能性がある物体のエージェント内でのみ発行される．

(5) 被支持物体認識要求 (to-be-supported recognition request)
支持可能要素を持った新しい候補が生成された時に，この認識要求が発行される．新候補の支持可能要素とその周辺領域が認識対象となり，適切レベルで再認識が行われる．この認識要求は，関係知識に基づいて，新候補に支持される可能性がある物体のエージェント内でのみ発行される．

(6) 未認識領域認識要求 (recognition request for vacant regions)
認識処理の最後の段階で，画像中に候補が検出されていない領域があれば，その未認識領域に対して，再認識を行う．

6.4.2 システムの動作の概要

第4章で述べたシステムと同様に，エージェントおよびそれを構成するモジュール，オブジェクトの動作は，すべてメッセージ駆動によって行なわれる．動作の概要は，第4章で述べたシステムとほぼ同一であるが，新たに導入した候補オブジェクトに関する部分は異なるので，一部重複するが説明する．ここでは，図6.6の例に従って，メッセージと認識要求の流れの詳細について述べる．

(a) 初期認識
初めに通信モジュールから「**初期認識要求**」が認識モジュールに送られる (図 6.6(a))．初めに「初期認識要求」を送信する時は，必ず最も解像度の低いレベル 0 を指定する．もしそれで，すべてのエージェントの認識モジュールが物体候補を検出できなかった場合は，各認識モジュールは 1 つずつレベル上げて「初期認識要求」を送信することになる．

(b) 候補オブジェクトの生成
認識要求によって認識モジュールが認識を開始し，1 つ物体を認識する度にその形状評価値を含む候補情報を通信モジュールに送る (b1)．それを受け取った通信モジュールは，支持関係，関係知識の評価を行い，関係評価値を求めて，他の物体の候補情報と照合し，競合がないかあっても自分の物体候補の方が正しいと判断できる場合は，候補オブジェクトを生成する (b2)．その後，新たに生成された候補オブジェクトは，他のすべてのエージェントの通信

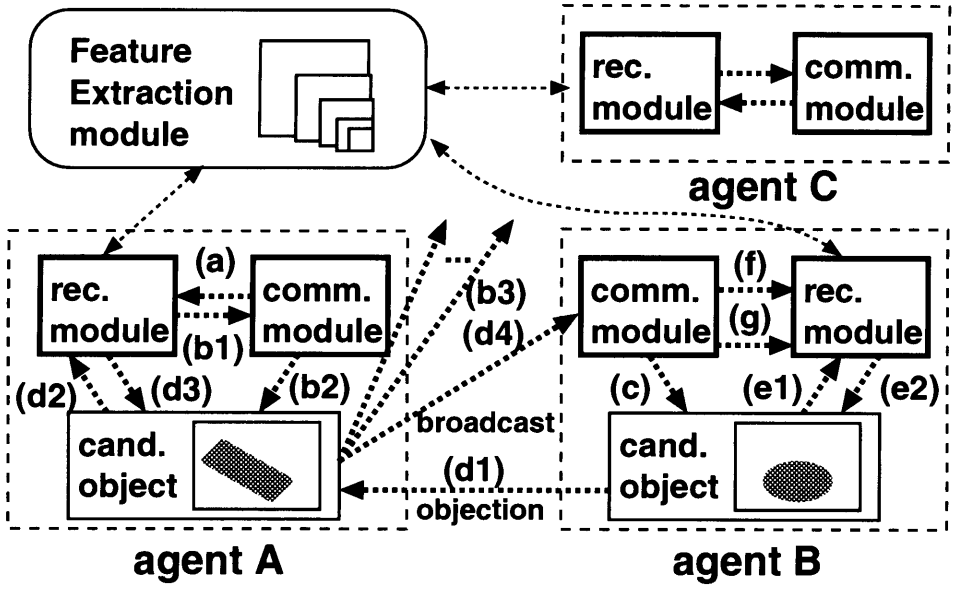

図 6.6 メッセージの流れ．(a) **初期認識要求**．(b1) 物体候補の情報．(b2) 候補オブジェクトの生成．(b3) 候補情報のブロードキャスト．(c) 関係候補発生の通知．(d1) 異義メッセージ．(d2) **候補変形要求**．(d3) 変形された物体候補の情報．(d4) 取消しメッセージ．(e1) **候補更新要求**．(e2) 更新された物体候補の情報．(f) **支持物体認識要求** もしくは **被支持物体認識要求**．(g) **未認識領域認識要求**．

モジュールに自分自身の候補情報をブロードキャストする (b3)．ここで，一度他の物体候補と競合をチェックしている理由は，できるだけ無駄なエージェントの生成と，候補情報のブロードキャストを起こさないためである．

(c) 候補情報の受信

ブロードキャストされた候補情報を受け取った他のエージェントの通信モジュールは，それが自分のエージェントの候補領域と競合していないか，もしくは関係知識に基づく関係が成立していないかチェックする．もし競合や関係があれば，自エージェントの該当する候補オブジェクトにその旨を通知する (c)．

(d) 競合解消

競合の通知を受け取った候補オブジェクトは，競合相手の候補オブジェクトに直接，異義メッセージを送信し (d1)，物体候補の評価値の比較による競合解消の処理を行う．比較の結果，評価が高い方がそのまま残り，低い方は認識モジュールに「**候補変形要求**」を送信して競合が起こらないように自分自身の候補領域を変形させる (d2)(d3) か，それが不可能な場合はその候補自身の取消しになり，取消しメッセージが他のエージェントに対してブロードキャストされる (d4)．

(e) 候補オブジェクトによる候補領域の更新

候補オブジェクトは，もし候補が適切レベルか適切レベルより上位のレベルで生成されたのでなければ,「**候補更新要求**」を認識モジュールに出すことによって，適切レベルで同じ領域を再認識を行う (e1)(e2).

(f) 支持関係に基づく存在領域の推定

通信モジュールは受け取った候補情報の候補がどの物体からも支持されていない場合に，自分の担当する種類の物体との間に支持関係の存在が関係知識から推定される場合は，自分の物体の存在可能性の高い領域を推測し，認識モジュールに対して「**支持物体認識要求**」を送信する (f). また，受け取った候補情報の候補に支持される可能性が関係知識から推定される場合は,「**被支持物体認識要求**」を送信する (f).

(g) 未認識領域に対する再認識 すべてのエージェントの認識が終了し，メッセージ待ち状態になった時に，ある一定の面積以上の未認識の領域が存在したら，その領域に対してより高い解像度の画像を用いて「**未認識領域認識要求**」を行う. 未認識領域が一定面積以下であれば，システム全体の認識が終了したとみなす.

(h) 候補の復活

再認識などによって評価値が変動して競合解消の結果が逆転する場合や，競合相手の候補が別の候補に取り消された場合，一度取り消された候補を復活させ，常に互いに整合性のとれた認識結果のみを残すようにする.

6.5 実験

比較的単純な室内シーンの画像を対象とするプロトタイプシステムとして,「机」「椅子」「ワークステーション (以下 WS と略す)」「本」「ペン」「床」の 6 種類のエージェントを構築し，6 台の PC (Intel Celeron 450MHz,128MB) からなる PC クラスタ上に PVM [114] を用いて，1 PC 1 エージェントとして実装した.

6.5.1 動作例

サンプル画像 1 (図 6.7, 1280 × 960) は「机」の上にディスプレイとキーボードからなる「WS」と,「本」「ペン」が存在しているシーンを表している．このシーンに対して，予め縮小した画像 (320 × 240) を用いて単一の解像度のみで認識を行った結果が図 6.8である．机，床，WS が認識できているが，ペンや本は認識できていない.

一方，複数解像度を用いた実験では，画像縮小率 0.7 として 1280 × 960 から 308 × 231 までの 5 段階 (レベル 0〜4) の画像を用いた．サンプル画像 1 に対しては，初めのレベル 4 に対する初期認識で，机，床，WS の候補が生成された．次に `on(book,desk)`, `on(pen,desk)` の関係知識によって，机領域とその周辺部を「本」エージェントと「ペン」エージェントがレベル 3 の画像を使って「被支持認識要求」による再認識し，図 6.10に示すように，本，ペンの候補が生成された．両方の候補とも，その後「候補更新要求」によってレベル 0 の画像を用いて再認識され，正しい結果が得られた．図 6.11はペンに対する候補更新要求時にレベル 0 で指定された領域である．最終的には，結果は図 6.9の様になった．なお，この結果が得られるまでに，WS のキーボード部分と本の競合や，左手の戸棚の一部が本やペンと候補として認識されることがあったが，それぞれ競合解消処理や支持関係のチェックによって最終的には取り消されている.

表 6.3はサンプル画像 1(図 6.7, 1280 × 960) と，320 × 240, 640 × 480 の大きさにそれぞれ縮小した画像を単一解像度で実行した時の実行時間と，複数解像度を用いた時の実行時間を示したものである．単一解像度の場合，320 × 240 の時には本とペンが認識出来なかったが，640 × 480 を用い

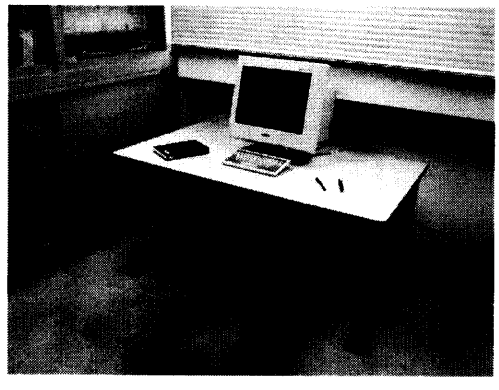

図 6.7 室内サンプル画像 1.

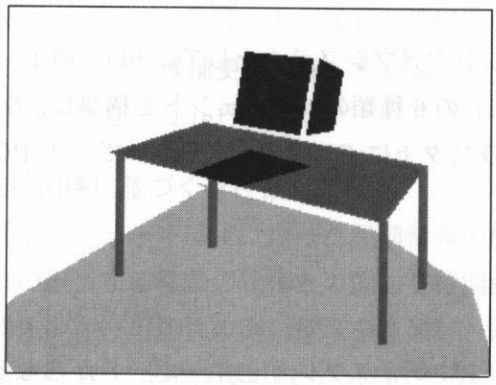

図 6.8 320 × 240 の単一画像を用いた時の認識結果.

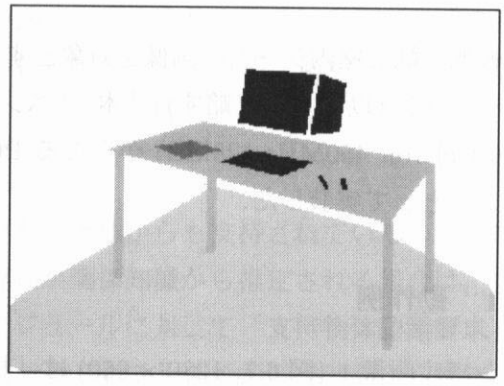

図 6.9 複数解像度を用いた認識結果 1.

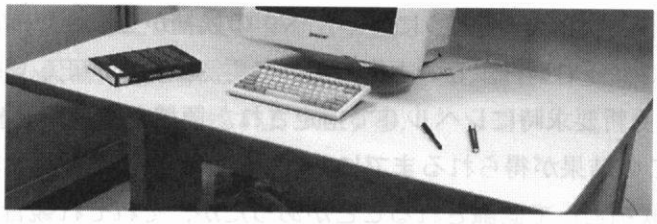

図 6.10 机上領域の拡大.

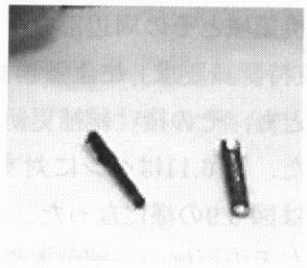

図 6.11 ペン候補領域の拡大.

た時はすべての物体が認識出来た．けれども，実行時間は約 5 倍にもなってしまっている．また，1280 × 960 の場合はメモリエラーによって実行できなかった．一方，解像度選択を利用することによって，元の画像が 1280 × 960 であっても，320 × 240 の 2.5 倍程度の時間で実行できている．

サンプル画像 2 (図 6.12, 1280 × 960) についてはサンプル画像 1 に比べてやや複雑となっていて，4 台の WS が奥の机の上に並んでいる部分のシーンの構造が従来の 320 × 240 程度の縮小画像では認識不可能であった．ここでは，5 レベルの解像度を必要に応じて使い分ける事によって図 6.13の様に WS，机，床などがほぼ認識ができている．

サンプル画像 3 (図 6.14, 1280 × 960) は，サンプル画像 2 と同様に，比較的複雑な画像である．認識結果は図 6.15に示すように，後方の WS が 1 台認識出来ていないものの，他の 5 台はうまく認識出来ている．特に，後方 3 台の WS は画像中では小さな領域としてしか現われていないのにもかかわらず，多重解像度の利用によって，うまく認識が出来ている．

表 6.3 実行時間の比較.

解像度	実行時間 [秒]
320 × 240	8.3
640 × 480	38.3
1280 × 960	× (out of memory error)
multi-resolution	19.6

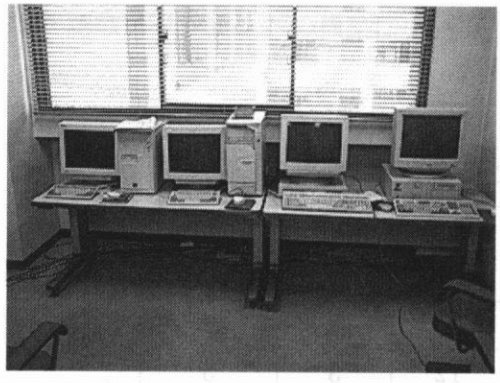

図 6.12 室内サンプル画像 2.

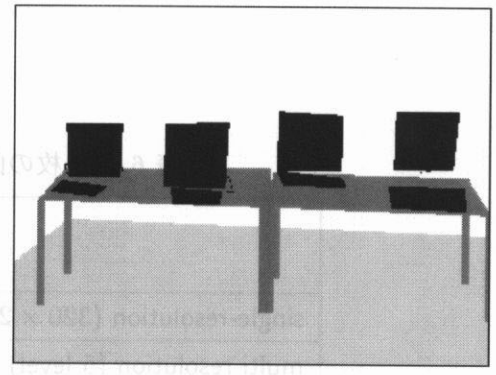

図 6.13 認識結果 2.

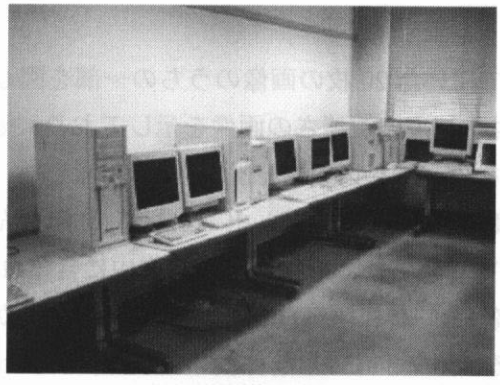

図 6.14 室内サンプル画像 3.

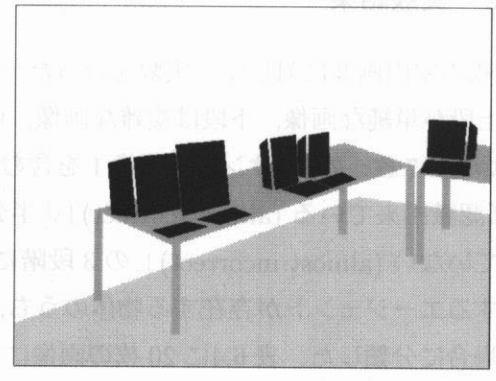

図 6.15 認識結果 3.

図 6.16 評価に用いた画像の一部．上段が単純な画像，下段は複雑な画像，中段は中程度の複雑の画像．

表 6.4 20 枚の画像に対する認識結果．

	almost correct	half correct	almost incorrect
single-resolution (320 × 240)	5	8	7
multi-resolution (5 level)	12	3	5

6.5.2 実験結果

20 枚の室内画像に対して，実験を行った．実験で用いた 20 枚の画像のうちの一部を図 6.16に示す．上段が単純な画像，下段は複雑な画像，中段は中程度の複雑さの画像を示しており，それぞれについて，7 枚，7 枚 (サンプル画像 1 を含む)，6 枚 (サンプル画像 2, 3 を含む) 用意した．結果は「ほぼ認識出来ている (almost correct)」「半分程度認識出来ている (half correct)」「ほとんど認識出来ていない (almost incorrect)」の 3 段階に分けて評価した．それぞれ，画像中に含まれていて，対応するエージェントが存在する物体のうち，80%〜100%, 30%〜80%，0%〜30%の物体が認識された場合に分類した．表 6.4に 20 枚の画像に対する 320 × 240 の単解像度の場合と，1280 × 960 の画像を多重解像度解析を用いた場合の実験結果を示す．

単解像度の場合は almost correct は 5 枚のみであった．しかし，多重解像度の場合は 12 枚の画像が almost correct となった．これは，多寿解像度解析を導入したことによる性能向上の表れである．一方，almost incorrect はそれぞれ 7 枚，5 枚で大きな差は無かった．これは，画像があまり

に複雑過ぎて，初期認識において有意な画像特徴をほとんど抽出できなかったために，効果的な再認識が出来なかったためである．

6.6 まとめ

本章では，我々が提案しているマルチエージェントによる画像理解システムにおいて，解像度選択による多重解像度解析を実現し，処理時間をあまり増大させることなく，効率的に高解像度の画像を認識に利用する方法を提案し，プロトタイプシステムとして実装し，実験を行った．

今後の課題としては，アルゴリズムレベルでの多重解像度解析の利用や，定性的モデル当てはめの精度向上などがあげられる．候補オブジェクト同士の動的なインラクションによる候補領域変形のメカニズムを充実させることも考えられる．他には，多数のエージェントの実装，多数の画像を用いたシステムの性能評価などが挙げられる．ただし，現状では，根拠特徴の抽出を行う認識モジュールは人手 (hand-coding) によって作成しているため，多様な状況に対応する認識モジュールを構築することは困難であり，特に人手によるモデル化が困難な複雑な形状の物体についての認識モジュールの構築や，多種類の認識モジュールの構築は容易ではない．そこで，次章以降では，学習を用いた認識エージェントの構築について検討を行う．

第 7 章

多数の学習画像を用いた画像認識

7.1 はじめに

本章では，学習を用いた実世界画像の classification 的な認識について従来の研究をまとめ，特に画像検索手法を用いた画像分類の研究について詳しく述べる．そして，従来の研究では，学習画像を収集することが困難であったために，顔画像や自動車の画像などに対象が限定されていた，もしくは特定の画像コレクションを利用した実験しか行われていなかったという問題点を指摘する．

7.2 学習による画像認識

第 4章，第 5章，第 6章では，マルチエージェントによって，多数の画像認識モジュールを統合してシステムを構築する方法について述べた．この方法では，認識モジュールの構築方法を特に規定しておらず，前章までの実験では，認識モジュールはすべて人手で構築 (hand-coding) していた．しかしながら，実世界の物体は高い視覚的多様性を持っているため，人手によって柔軟な認識を行うことのできる認識モジュールを構築することは困難な作業であり，したがって，多くの種類の認識モジュールを構築することは事実上不可能であった．「机」「椅子」の様な比較的簡単な形状の人工物の場合は，人手による構築でもある程度の性能を持った認識モジュールを構築することが可能であるが，人工物以外の物体や，複雑な形状の人工物を対象とする認識モジュールは人手では難しく，学習によって構築するしかない．そこで，本章以下では，実世界に存在する物体を柔軟に認識することができる認識モジュールを学習によって自動的に構築することを考える．

学習による認識では，認識を行うための準備である学習を行う学習フェーズと，認識を行うための認識フェーズが存在する．初めに学習フェーズで，画像と望ましい認識結果の組みである学習データをシステムに与えることによって，画像知識ベースを自動的に構築し，次に認識フェーズで，入力画像に対して蓄積した知識を利用して認識を行い，結果を出力する．

第 2章で述べたように，学習を用いた認識には大きく分けて 2 通りの方法が存在する．1 つは，予め対象に応じた特徴抽出方法を用意しておいて，それを用いて学習画像から画像特徴を抽出してモデルを構築し，認識対象画像から抽出した画像特徴と比較することによって，クラス分類を行うパ

ターン認識手法による方法．もう 1 つは，認識対象画像とその画像中の認識したい要素やその特徴を併せて入力すると，自動的にその認識の手順を生成する画像処理エキスパートシステムである．後者の画像エキスパートシステムについては，第 2章で述べたように，複雑な形状の物体や 3 次元的な物体の抽出を行うことは処理手順の組合せ爆発のために難しく，1980 年代後半に盛んに研究されたものの現在はあまり研究は盛んに行われていない．そのため，本章では，画像エキスパートシステムについては取り扱わず，前者のパターン認識手法による方法をのみを扱うこととする．

パターン認識は，画像に限らず実世界の様々なパターン情報，例えば，文字，音声，生物の行動パターンや DNA 配列などを，予め定められた複数のクラス (class) のうちの一つに分類する研究分野であり，統計理論にその基礎を置いている [39, 40]．パターン認識の研究は，画像認識の研究が始まる以前から行われており，現在でも盛んに研究が行われている．まさに「認識」の基礎といえる研究分野である．パターン認識の研究においては，各パターンデータから特徴抽出を行った後の特徴ベクトル (feature vector) とクラスの分類の関係から，入力パターンをクラス分類するための識別器 (classifier) を求めることが研究の中心であり，特徴抽出の方法については個別の分野で研究されている．特徴抽出では，認識対象の特徴をよく表す，つまり他の対象との違いを良く表す特徴をパターンデータ (画像認識の場合は画像) から取出す必要があり，基本的には対象に依存したものとなる．そのため，画像認識において，パターン認識手法を適用するためには，やはり対象を想定して特徴抽出を行う必要があり，現状では，学習を用いた classification においては，ほとんどの研究において認識対象が限定されている．主に応用範囲が広い対象が研究され，顔画像 [41, 42, 43, 44] や自動車 [44, 61, 62] などの検出がよく研究されている．

そうした中でも，近年，対象に依存しない画像特徴を利用して，大量の画像を学習画像として画像認識を行う example-based による一般物体認識 (generic object recognition) の研究が行われるようになっている．例えば，一般的な特徴である画像中の領域の配置の関係を学習して，認識したい各クラスについてテンプレートを自動構築するという A. L. Ratan らによる研究 [118] がある．この研究は，従来より行われていた人手によって構築されたテンプレートを利用した認識 [119, 120] を拡張したものである．ただし，自然画像を対象として「雪山」「山と湖」「原野」「滝」などを分類した結果，人手によるテンプレートを用いた場合の認識率は 6〜7 割程度である一方，学習によって自動構築した場合は 2〜3 割程度の認識率しか得られていない．J. R. Smith らによる研究 [121] でも同様に領域の位置関係をテンプレートとして学習しているが，こちらは 10 クラスの画像分類について学習画像 91 枚，テスト画像 266 枚で実験を行い，70.7%の精度で分類を行っている．だだし，同一のデータセットに対して通常のカラーヒストグラムを用いて分類実験を行った結果，67.3%の精度であったので，学習画像，テスト画像ともに同一クラスの画像は互いに類似した画像が多かったと考えられる．

また，M. Weber らの研究 [61, 62] では，ある特定の種類の物体を含む画像と含まない画像を数百枚ずつ用意して，それから自動的に物体の特徴を学習することができる．認識対象は小領域の集合体として表現され，領域のパターンとその位置関係によって認識が行われる．O. Maron らの研究 [122, 123] でも，同様に正サンプルと負サンプルを用いて，正サンプル画像に共通に含まれていて，負サンプル画像に含まれない画像特徴を統計処理によって発見し，それを学習することが出来

る．また，H. Schneiderman らの研究 [44] では，対象に依存しない特徴であるウェーブレット変換係数を画像特徴として利用することによって，同じシステムで顔画像と乗用車の両方の画像が認識可能となっている．

一方，近年，コンピュータの進歩により大量の画像を高速に処理できるようになってきた．そこで，画像データベースおける検索手法を物体認識 (分類) に応用するという試みがなされている [55, 56, 57]．画像内容に基づく画像検索 (content-based image retrieval, CBIR)[124, 125]，つまり，画像特徴を手がかりとした画像データベースの検索では，質問画像が与えられて，それに類似した画像がデータベースから画像特徴に基づいて検索される．なお，“content-based” と言うものの，実際には画像特徴に基づく (image-feature-based) という意味であり，画像の意味内容に基づく (こちらは，semantics-based ということもある) という意味ではなく，画像特徴を用いない keyword-based (もしくは meta-data-based) の対義語の意味でしかない．画像検索においては通常の画像認識とは異なり，データベース中に様々なジャンルの画像が含まれるので，予め画像の種類を限定することが出来ない．そのため，どのような種類の画像でも対応可能な色情報やテクスチャ情報が画像特徴として用いられる．画像検索の手法を用いた画像認識 (分類) では，基本的には含まれている主な物体の名称が予め分かっている学習画像を大量に用意して，認識したい画像に類似している学習画像を検索し，その学習画像に含まれている物体が認識対象画像にも含まれていると見なして，認識 (分類) を行ったこととする．こうした方法は，画像同士の類似度の計算方法を定めて，学習画像を大量に集めれば，認識が可能となるので，比較容易にシステムが実現できる反面，すべての対象に対して適用可能な一般的な方法を用いているために，認識 (分類) 精度を高くすることが難しいという問題がある．なお，画像検索の手法を用いた認識では，画像中からの認識対象の切出し (segmentation) を行わないことが普通なので,「画像認識」ではなく「画像分類」と呼ぶ方が一般的である．以上の画像検索手法による画像分類の研究については，次節で詳しく述べる．

7.3 画像検索手法による画像分類

画像データベースの分野において，画像特徴に基づく画像の検索や分類が，内容に基づく画像検索 (content-based image retrieval, CBIR) として従来より研究されている．キーワードを用いた画像検索 (keyword-based image retrieval) では，人手によって予めすべての画像データにキーワードを付けておく必要があるが，CBIR では各画像データから画像特徴を自動抽出して画像間の類似度を判定し，ユーザが指定した画像に類似した画像を検索する．CBIR では，画像が表すシーンの意味内容を考慮することはせずに，画像の表層的な特徴量の類似性によって，画像の類似性を判定することを行っている．

形容詞を中心とした印象語や感性語を検索キーワードとして，予めキーワードと画像特徴の対応付けを心理実験から統計的に求めておくことによって，画像を検索する手法も提案されている [126, 127]．画像全体の印象は，画像全体の色の分布などに大きな影響を受けるので，これらの研究では画像の意味内容を考慮していないのにも拘らず，一定の成功を納めている．

画像特徴を用いた画像検索の手法を用いて，画像の意味内容による分類を目指した研究としては，

最も古典的な研究として，画像をブロック部分領域に機械的に分割して，それぞれの部分領域の特徴量と名詞単語の関連付けを行った Photobook [128] の研究がある．この研究では，事前の学習において，ユーザが領域と単語の対応を指示してやる必要があった．

一方，森らの研究 [57] では，百科事典中の画像と説明文から部分領域と単語の対応を自動的に学習する方法を提案している．この研究では，1 つの画像に複数個の単語を持たせて，学習画像の部分領域を特徴量に関してクラスタリングし，各クラスタについて各単語の出現確率を予め求めておく．そして，テスト画像の各部分領域について，最も近いクラスタの単語出現確率の平均値の上位の単語がテスト画像の関連単語ということとしている．しかし，百科事典の扱う対象があまりにも広範囲に渡っているために，良好な精度が得られているとは言い難い．同じ手法を WWW から収集したテキストと画像に対して行った研究 [129] もある．

C. Y. Fung ら [130] も同様にクラス既知の画像をブロック分割し，次にすべての学習画像のブロックをクラスタリングする．そして，学習画像をそのブロックが分類された先のクラスタの組合せによって表現することとし，各クラスの平均的なクラスタの組合せを求める．このクラスタの組合せによる表現のことを picture words と呼んでいる．次に未知画像の各ブロックをクラスタに分類し，その組合せからクラスを決定する．

同様な研究で，単純なブロック分割ではなく，カラー領域分割アルゴリズムを用いて画像を分割して，各領域の特徴量に基づく類似特徴検索による画像認識が提案されている．S. Belongie らによる研究 [55, 131] では，Blobworld[132, 133] と呼ばれる領域分割表現を用いて，各領域の特徴量に基づく類似特徴検索による画像認識が試みられている．また，K. Barnard らによる研究 [56] では Blobworld を用いて，[57] と類似した方法で，領域と単語の対応付けを行っている．この 2 つの研究では，同一クラスの画像が互いに類似しているものが多い Corel 社の画像コレクションが評価に用いられているために，前者が分類の精度が 5〜6 割程度，後者が上位 15 個の単語の内に画像に関連する単語が 8 割程度の確率で含まれるという，比較的よい精度が得られている．また，同様の手法を用いて，芸術品の画像分類した研究 [134] もある．[56] と同じ研究グループの P. Duygulu らによる [135] では，normalized cuts [136] による領域分割を用いて，画像全体ではなく，領域分割された領域毎にラベル付けを行っている．1980 年代に盛んに行われた領域分割とラベリングによる画像認識とは，大量の学習画像から学習する点で大きく異なっている．学習データは，画像とその画像中に含まれる複数の物体の名前である．画像中の領域と与えられる物体名の対応は学習時には与えられることはなく，統計処理によってシステムが自動的に推定する．

A. B. Benitez らによる研究 [137, 138] では，テキストによる説明文の付いた画像を対象に，文の中の単語と画像特徴との対応付けをクラスタリングを用いて行っている．対応付けを行う際に電子的な単語辞書である WordNet[139] を利用して，単語同士の意味的な近さも考慮している．一方，R. Zhao ら [140, 141] は WordNet のような辞書を利用するのではなく，潜在意味的インデキシング法 (latent semantic indexing, LSI)[142] を用いて説明文中の単語の共起性に基づいて，単語同士に意味的な類似性を考慮することを提案している．

J. Huang らの研究 [143] は画像を 2 分木によって階層的に分類する方法で画像分類を行っている．分類木の各ノードで，それぞれ特異値分解 (singular value decomposition, SVD) を行って，それぞ

れ異なる特徴ベクトルの圧縮を行っているのが特徴である．

なお，これらの研究では，様々な種類の画像を認識対象とすることできるために，顔画像や自動車画像の様に対象が限定された場合と違って，研究者自身が独自に実験に用いる画像を収集することが極めて困難である．そのため学習画像として，この種の研究の事実上の標準評価画像となっている 6 万枚の著作権フリーの画像を含んでいる Corel 社の Corel Image Library が使われている場合が大部分である．Corel Image Library は枚数は多いものの，様々な画像が含まれているので，同一のカテゴリーに含まれる画像は多くても 100 枚程度に過ぎない．しかも，プロの写真家が撮影した整った画像のみを集めていて，同じカテゴリに含まれる画像は同一のカメラマンが似たような構図で撮影した場合が多く，必ずしも実世界画像に対する認識システムの評価に適した画像であるとは言えないという問題がある [14].

また，画像を具体的なクラスに分類するのは困難で，なかなか実用的な精度を得られないので，その代りに街中の画像と遠景画像 (city–landscape classification)[144, 145]，室内画像と屋外画像 (indoor–outdoor classification)[146]，写真画像とコンピュータグラフィクス画像 (CG 画像)(photographics)[147] のように大まかな分類を行う研究がある．これらの研究ではテスト画像からそれぞれのクラスの特徴を学習し，9 割程度の実用的な分類精度を誇っている．Q. Iqbal ら [148, 149] は，画像から抽出したエッジ画像に対して画像認識の手法である perceptual grouping[64] の手法を用いてを L 交点，U 交点などをグループ化し抽出し，それらの数を特徴量として，風景画像について建物が含まれるかどうかを判定した．分類精度は 7〜8 割程度であった．一方，A. Vailiya ら [150, 151] は以上述べたような 2 クラス分類の分類器を複数個階層的に組合せることによって，画像を indoor, city, sunset, forest, mountain に分類することを実現し，9 割程度という高い分類精度を実現した．同様に，A. Hartmann と R. Lienhart[152, 153] は Web から収集した画像に対して，写真と CG 画像，プレゼンテーションスライド画像，コミック画像について階層的に分類し，9 割以上の分類精度を実現した．

L. Wenyin らの研究 [154] では，最初から完全な分類を目指すのではなく，ユーザからのフィードバックを利用して，逐次的に学習し分類精度を改善していく半自動画像分類システムが提案されている．

7.4 まとめ

本章では，学習による classification 的な画像認識についてその従来研究をまとめた．Identification のための画像学習方法は数多く提案されているものの，それに比べると classification は顔画像などごく一部の応用範囲の広い対象を除くと多くはない．それは，特徴抽出が対象に依存した方法でないと，認識精度が低くなってしまい，実用的な認識とすることが難しいからである．しかし，そうした問題があるにもかかわらず，近年，計算機の性能向上による大量の画像の処理が可能になったことから，大量の画像を学習画像とする認識の研究が行われるようになってきている．対象を限定しない場合，CBIR の手法を応用する場合が多く，まだ実用的な認識精度になっているとは言えないが今後の研究は多いに期待される．

また，従来の研究では，学習画像を収集することが困難であったために，顔画像や自動車の画像などに対象が限定されていた，もしくは特定の画像コレクションを利用した実験しか行われていなかったという問題点があった．そこで，次の第 8章では幅広い種類の画像を WWW(World-Wide Web) から自動収集する方法について述べ，さらに第 9章では，その収集画像を用いた画像分類の実験について述べる．

nnn