一般物体認識の現状と今後

柳井啓司†

「一般物体認識」とは，制約のない実世界シーンの画像に対して計算機がその中に含まれる物体を一般的な名称で認識することで，コンピュータビジョンの究極の研究課題の 1 つである．人間は数万種類の対象を認識可能であるといわれるが，計算機にとっては，同一クラスに属する対象のアピアランスが大きく変化するために以前はわずか 1 種類の対象を認識することすら困難であった．ここ数年，新しいモデル表現の提案，機械学習法の進歩，計算機の高速化などにより，急速に研究が進展しており，現在は 101 種類の対象に対して 6 割程度の精度で認識が可能となってきている．本論文では，一般物体認識研究のサーベイを手法に加えて，データセット，評価ベンチマークについて行い，さらにその今後について展望する．

The Current State and Future Directions on Generic Object Recognition

Keiji Yanai†

"Generic object recognition" aims at enabling a computer to recognize objects in images with their category names, which is one of the ultimate goals of computer vision research. The categories which are treated with in generic object recognition have broad variability regarding their appearance, which makes the problem very tough. Although human can recognize ten thousands of kinds of objects, it is extremely difficult for a computer to recognize even one kind of objects. For these several years, due to proposal of novel representation of visual models, progress of machine learning methods, and speeding-up of computers, research on generic object recognition has progressed greatly. According to the best result, the 66.23% precision for 101-class generic image recognition has been obtained so far. In this paper, we survey the current state of generic object recognition research in terms of datasets and evaluation benchmarks as well as methods, and discuss its future directions.

1. はじめに

1.1 一般画像認識とは？

今日，我々の日常にはデジタル化された写真が大量に存在している．それらのデジタル写真は様々な実世界シーンの "一般的な" 画像であり，従来の画像認識の研究で対象としてきた特定の制約の下で撮影された画像とは大きく異なる．そうした制約のない実世界シーンの画像に対して，計算機がその中に含まれる物体を一般的な名称で認識することは **一般物体認識（generic object recognition）** と呼ばれ，画像認識の研究において最も困難な課題の 1 つであると考えられている．なぜなら，制約のない画像における「一般的な名称」が表す同一クラスの範囲が広く，同一クラスに属する対象のアピアランスの変化がきわめて大きいために，(1) 対象の特徴抽出，(2) 認識モデルの構築，(3) 学習データセットの構築，が困難なためである．

画像と言語を対応付けることを目指した一般物体認識は，画像認識の研究が始まった今から 40 年以上前より研究が行われている．しかしながら，いまだに人間の顔の正面画像を除いては，実用的な精度で認識が可能な対象はほとんどない．人間は数万種類の対象を認識可能であるといわれている[1]．その一方で，「山」「椅子」「ラーメン」などの我々にとって馴染み深い対象が写った画像を計算機によって自動的に検出することは，現状ではきわめて困難である．そうした状況ではあるが，ここ最近の 5 年間，一般物体認識はそれ以前とは違って急速に進歩をとげてきている．これは，part-based と呼ばれる新しい物体表現法の登場，機械学習手法の進歩，計算機の高速大容量化によるところが大きい．

一般に実世界画像を対象とした物体認識には大きく分けて identification（同定）と classification（分類）の 2 種類の認識がある[2]．Identification は既知の物体

† 電気通信大学情報工学科
Department of Computer Science, The University of Electro-Communications

を画像から見つけ出す認識であり，入力画像とデータベース中のモデルの照合を行い，どのモデルに対応する物体が画像中に存在するかどうかが出力となる．一方，classification は画像中の未知の物体をそれが属するべき既知のクラスに分類する認識であり，画像中の物体とクラスとの対応付けの結果が出力となる．「物体認識」というと identification のことを指すのが一般的であるが，「一般物体認識」は classification を意味する．

このように一般物体認識は classification としてのパターン認識問題であるという側面を持つが，それに加えて画像と言語の対応付けという目的を持っている点が通常のパターン認識問題とは異なっている．そのため，一般物体認識においては学習データセットのクラスラベルは単なる記号ではなく，物体の一般的な名称と対応している必要がある．しかしながら，最初に述べたように，一般名称が表す同一クラスの範囲がきわめて広いために，クラスと一般名称の対応がとれた学習データセットを作成することはそれ自体困難な問題である．逆に，クラスと一般名称の対応が十分にとれたデータセットが存在するとすれば，クラス内の多様性のためにそれに対するパターン認識問題がきわめて難しい問題となってしまい，認識技術の方が対応困難になることが予想される．そのため，現状の一般物体認識の研究では，問題が難しくなりすぎないように，たとえば「バイク」「自転車」は真横から見た画像のみというような適度な限定を加えて人手により構築された学習データセットが共通データセットとして広く利用されている．

一般物体認識においては，かつては何をどの程度認識すればよいのかが曖昧であったが，現在では共通データセットがそれを明確化している．共通データセットを学習および評価データとする認識において高い認識率を出すことが今日の大部分の一般物体認識研究の目的となり，クラスと一般名称の関係はデータセット任せになっている．それにともなって，特定の共通データセットという枠の中で良い性能を発揮する手法が，そのまま Web 上などに存在している一般の画像に対しても有効であるのかという疑問も生まれてきている．そこで近年は，認識の技術レベルが向上するにつれて，データセットに対する限定が弱められる傾向にあり，一般物体認識のための共通データセットは，物体の一般名称に対応するような様々な画像サンプルを含むように徐々に多様かつ大規模なものに進化してきている．実際，2006 年までは 101 種類 9,144 枚のデータセットが標準的に用いられていたが，2007 年になってからは 256 種類 30,607 枚のデータセットが登場した．現状はまだ発展途上の段階にあるといえるが，今後はさらに大規模化することによって，一般物体認識研究用の共通データセットがクラスと一般名称の対応がとれた一般的な画像データセットに近付き，提案される手法もより一般的な画像に対応した手法に進化してゆくことが期待される．

現状の一般物体認識研究においては「顔」「自動車」「ライオン」など様々な対象に対して，学習データを替えるだけで検出器を自動的に構築できる一般性のある認識手法の研究に重点が置かれている．つまり，現在の一般物体認識の手法は，トレーニングデータからの学習が行えることが必須で，かつてさかんに行われた人手によって作り込まれたプログラムやルールのみによって認識を行う方法とは一線を画している．仮に単一の対象に特化した○○検出器をあらゆる対象について人手で作り出し，それらを並列に動作させることができれば，それも一般物体認識の実現法の 1 つであるということができるだろう．しかしながら，実際には扱うべき対象の種類が膨大であるために，そのような実現法は現実的ではなく，学習データの交換によって様々な対象に適用可能な一般的な方法が求められている．

なお，一般物体認識という場合に，形のある「物体」のみを認識対象とする場合もあるが，本論文では直接対応する物体がない「夕焼け」「海岸」「運動会」などの「シーン」の認識も一般物体認識の一部に含めて考えることとする．さらに，広く考えれば，名詞以外の形容詞や動詞で表現される言語概念も認識対象とすることが可能であるが，現状では一般的な画像に対してそうした認識を行う研究は一般的ではないので，本論文では物体もしくはシーンを表す名詞概念を画像から認識することを，「一般物体認識」と呼ぶこととする．こうした認識を「一般画像認識」と呼ぶこともある．また，英語では，generic object recognition，generic image recognition 以外に，generic object categorization，category-level object recognition と呼ぶこともある．

一般物体認識では，本来は画像からの認識対象物体の切り出しを含むべきであるが，現状では画像全体についてクラス分類を行う一般物体認識の研究が主流を占めているため，本論文では，画像全体をクラスに分類するタスク，画像から対象物体領域を切り出すタスク，画像中の対象の存在位置を矩形によって示すタスクのいずれもを「一般物体認識」であると見なすこととする．

1.2 背　　景

近年，デジタルカメラの普及やハードディスクの大容量化によって，一般の個人が大量にデジタル画像を蓄積することができるようになった．しかしながら，

計算機が画像の意味を理解することができないため，画像の取扱いに関する計算機と人間のセマンティックギャップは狭まることはなく，現状では大量の画像データの分類や検索には人手の介入が不可欠である．一般物体認識はそうした視覚情報処理におけるセマンティックギャップの解消のための技術として，実現が期待されている．たとえば，画像に対する自動キーワード付けや，画像の意味内容による分類や検索などの実現が一般物体認識の実現によって可能となることが期待できる．また，一般物体認識は，**機械による人間の高次視覚機能の実現**というサイエンス的な観点からも興味深い研究であるといえる．

一般物体認識は，主に計算機の高速化と機械学習技術の進歩によって，近年，アメリカとヨーロッパを中心にさかんに研究されるようになってきている．実際，現在，一般物体認識ブームにあるといってよい．ICCV，ECCV と並んでコンピュータビジョンに関する最大の会議である CVPR では，2006 年には 12 のオーラルセッションのうち 3 つが認識関係で，そのうち一般物体認識に関係する発表は 8 件あった．ポスターセッションも 12 のうち 3 つが認識関係で，ポスター論文にも多くの一般物体認識に関する論文があった．2 年前の 2004 年の CVPR では，認識に関するオーラルセッションは 1 つのみで，そのうちわずか 1 件のみが一般物体認識に関する発表であったことを考えると，これはまさにブーム到来ということができる．同様の傾向は ECCV でも見られる．しかしながら，日本国内ではほとんど研究が行われていない．そこで，本論文では，一般物体認識に関して，その研究の歴史，現状，今後の課題についてまとめ，国内研究者向けに紹介を行う．特に，現状に関しては，最新研究のサーベイに加えて，データセット，評価ベンチマークについても解説する．

本論文の構成を簡単に述べる．2 章では，従来の一般物体認識の研究について 1990 年代以前と 1990 年代に分けて解説する．3 章では，近年の一般物体認識ブームの先駆けとなった，統計的学習手法を用いた 2 つの方法について取り上げる．特に 2 つ目の方法は局所特徴量という新しい特徴表現を提案しており，これが一般物体認識における 1 つのブレークスルーとなったといえる．4 章では，最新の研究トピックについて解説する．局所特徴量を簡単に利用する方法である bag-of-keypoints，局所特徴量にサポートベクターマシンを組み合わせた研究，物体どうしの空間コンテキストを利用する研究などについて紹介する．5 章では，データセットとベンチマークワークショップについて述べる．6 章では，主に多クラス化とクラス内変化の対応に関して，今後の課題を議論し，最後に 7 章でまとめを述べる．

なお，本論文は 2006 年 9 月の CVIM 研究会の予稿[3] を加筆修正したもので，サーベイがカバーする範囲は 2006 年末までとなっている．2007 年以降についてはごく一部を除いて原則としてカバーしていないので，その点についてはご留意願いたい．

2. これまでの研究

本章では一般物体認識の歴史（**図 1**）について述べる．

2.1 1990 年代前半まで

一般物体認識もしくは一般画像認識は，画像認識の研究が始まった 1960 年代当初から研究が行われていた．しかし，当初より物体認識はとても困難な問題であることは認識されており，最初に成功をみた研究は，限定された世界『積木の世界』を対象としたものであった．その代表例ともいえる，2 次元平面上に線で描かれた物体の 3 次元構造を推定する線画解釈[4] は多くの研究が行われたが，線画そのもの，もしくは容易に線画を得られる画像のみが対象となり，実世界の画像からいかに正しく線画を抽出するかに関しては問題が解決されることはなかった．

その後，実世界画像に対する研究として，2 次元的な取扱いのできる画像，たとえば，航空写真などのような画像に対する理解システムがさかんに研究されるようになった．認識の方法は領域分割の延長線上にあり，同じ対象を表している領域を切り出して，その形状や色，模様，領域間の関係などを手がかりにしてラベリングすることによって認識を実現していた．あらかじめ物体の完全な形状モデルが得られない場合の実世界シーンの認識は，古くは Tenenbaum[5] らの領域分割した領域に対する緩和法によるラベリングによる認識があるが，こうした方法は非常に単純な方法であり，複雑な画像に対しては有効ではなかった．その後は Ohta によるシーン理解システム[6]，The Schema System[7]，SIGMA[8] などの画像中の物体ごとに認識手法を用意する知識ベース型の画像理解システムが登場した．認識のためのモデルはルールとして表現されていたが，ルールは人手によって記述していたため，認識対象を増やすことが困難であるという問題，すなわち人工知能研究における「知識獲得のボトルネック」の問題があり，それを解決することはできなかった．

当時の研究のほとんどが 3 次元画像を航空写真と同じように 2 次元的な画像として取り扱っており，領域分割を行った後に，関係や構造の情報を利用してそれ

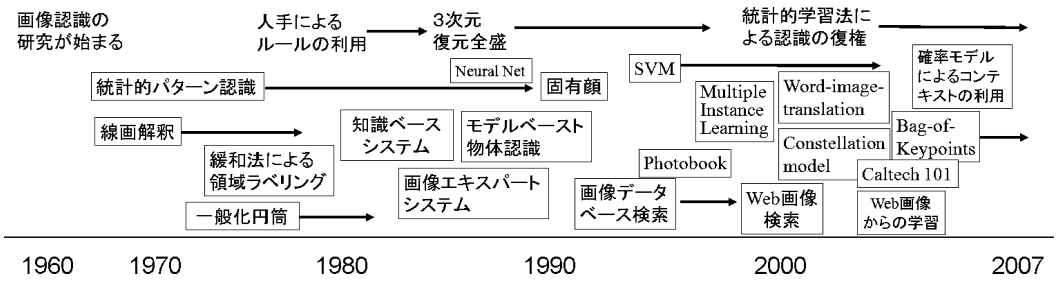

図 1 一般物体認識を中心とした画像認識研究の歴史
Fig. 1 History of research on generic object recognition.

ぞれの領域にラベリングを行い認識を実現していた．このような方法では，初期の領域分割の結果が最後まで結果に影響してくることや，対象が 3 次元であるにもかかわらず，3 次元的な取扱いがなされていないという問題点があった．そのため，その後，Marr の提案した「視覚認識への計算論的アプローチ」[9] の影響で，3 次元情報の復元が重視されるようになり，こうした領域分割＋ラベリング規則のような 2 次元的な物体認識の手法は下火となった[10]．

その後，3 次元の実世界を対象とする認識については，model-based による物体認識の研究がさかんに行われた[11]．Model-based による物体認識では，認識対象物体の形状モデルを知識としてあらかじめ用意しておいて，画像とモデルの照合を行うことにより，画像中の物体を認識する方法である．モデルの表現の最も一般的な方法は，物体の 3 次元幾何形状をモデルとするものである．ほかにも一般化円筒[12] を用いた構造表現によって対象を要素に分解してネットワークやグラフなどによって構造的に表現する方法や，パラメータによって形状モデルの形に幅を持たせることなども行われた[13]．1980 年代，90 年代においては「物体認識」という用語は，こうした identification を目的とした model-based による物体認識のことを一般に指していた．

これらの認識の方法は，どの表現方法も物体の形状を直接認識に利用していた．そのため，認識する対象の形状が完全に既知でないと，正しい認識が不可能であり，identification には向いているが，classification に適用することは困難であった．例外的に，プロトタイプモデルによって，model-based で classification を目指した研究[14] があったものの，実際に実世界の画像を認識しようとすると，実世界に存在する物体の形状は無限ともいえるほどあり，そのすべての形状が既知であることはありえず，また，海や道路などのように明確な形状を定義することができない物体も多く存在するなどの問題を解決することは不可能であった．

一方，違うアプローチからの手法も提案された．物体の機能を推測して機能から物体を認識する function-based recognition[15]，物体の候補を複数出して物体間の関係によって最終的な結果の選択を行う context-based recognition[16]，画像エキスパートシステム[17)~19)] などが提案されたが，結局ルールベースの認識手法には変わりなく，一般化することはできなかった．

2.2 1990 年代

1980 年代では人手によるルールや幾何形状モデルを認識モデルとして用いていたために認識対象を増やすことが困難であった．そこで，1990 年代では学習画像を用意して，それから自動的に特徴量を抽出し認識を行う研究が多く行われるようになった．

物体の形状を用いない方法として，テクスチャや色を用いる方法が提案された．Swain ら[20] によるカラーインデキシングはヒストグラムを用いる代表的な手法で，色の分布のヒストグラムを特徴量として類似画像の検索を実現した．見た目が類似している画像を高速に検索するのに向いているため，この手法は現在でも画像データベース検索の標準的な手法として用いられている．ヒストグラムを画像の一部分から作成することによって，大量画像データに対する部分画像検索にも応用されている[21),22)]．しかしながら，特徴量が色のみのため，簡単なシーンの認識以外の具体的な物体の認識には向いていない．ヒストグラムはテクスチャにも応用された[23]．

他に有力な手法として，濃淡画像の画素値をベクトルの要素と見なして，画像ベクトルを固有空間を用いて圧縮し，圧縮されたベクトルを特徴量と見なす固有顔[24] が Turk らによって提案された．この研究は顔画像に対する classification を目的としていたが，それを一般の 3 次元物体の identification に適用するパラメトリック固有空間法[25] が Murase らによって提案された．それ以前の 3 次元物体の identification では，

認識対象の 3 次元モデルをあらかじめ用意する必要があったが，パラメトリック固有空間法では，認識対象の 2 次元の外観（アピアランス）の視点位置に関する変化を固有空間中に多様体として表現することによって，3 次元物体の認識を実現した．この方法では，3 次元物体を，3 次元情報を復元せずに 2 次元のアピアランスのみで認識するので，appearance-based と呼ばれ，現在の物体認識の方法の基本的な考え方になっている．

これらの方法では，学習画像を用意すれば認識が可能となるが，認識対象の切り出しによって認識対象のみが写っている学習画像を用意する必要があり，種類を増やすことは容易ではなかった．また，認識対象全体を特徴として利用しているので，オクルージョンに対処できないという問題もあった．

1990 年代においては 3 次元の復元が重視されたため，classification を目的とした一般物体認識はあまりさかんに研究が行われたとはいえず，「一般物体認識の暗黒時代」であった．そうした中で，次に述べるように，画像検索研究から一般物体認識への強い要求が生まれつつあった．

2.3 画像検索からのアプローチ

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 1990 年代よりさかんに研究されていた[26),27)]．画像検索は画像認識のコミュニティではなく，主に，大量の画像を含む画像データベースや映像データを研究対象としてきたマルチメディアの研究コミュニティで研究されてきた．画像検索では，かつては見た目が類似している画像を検索することが主眼であったが，近年は意味的に類似している画像を検索することにその興味が移りつつある．意味的な類似とは，画像中の物体，もしくは画像の表すシーンのクラスが同じであるということである．もし，あらかじめクラスの分かっている画像を用意することができて，意味的な類似画像検索ができれば，それはまさに一般物体認識そのものである．つまり，「画像検索」と「物体認識」の求める方向が同じになったということができる．ただし，「画像検索」では大量の画像を対象としているので対象を限定せずに多くのクラスを扱うことが重視され，「物体認識」では種類の多さよりも単一もしくは少数のクラスの物体についての認識精度が重視される．

画像検索の手法を用いて，意味内容による画像の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（グリッド分割）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[28)] 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．

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．Belongie らによる研究[29),30)] では，Blobworld[31)] 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．この研究に確率モデルを導入したものが，次章で紹介する word-image-translation model である[32),33)]．

他に領域分割を用いる方法としては，自然シーン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 Ratan らによる研究[34)] 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シーン分類[35)] を拡張したものである．Smith らによる同様の研究[36)] もある．一方，Maron らの研究[37),38)] では，一般物体認識に multiple instance learning（MIL）[39)] を導入することを提案した．彼らは，正サンプルと負サンプルの学習画像データをグリッド分割し，正サンプル画像（positive bag）に共通に含まれていて，負サンプル画像（negative bag）に含まれない部分画像を，新たに diverse density という指標を導入することによって求め，未知画像が対象クラスであるかどうかの 2 クラス分類を行った．

こうした画像検索の発展系としての一般物体認識に関連する研究成果は，コンピュータビジョンの会議よりも，ACM Multimedia や ACM CIVR（International Conference on Image and Video Retrieval），IEEE ICME（International Conference on Multimedia and Expo）などのマルチメディア系の会議で発表されることが多い．

3. 21 世紀の新しい手法：統計的機械学習による方法

2000 年代になると，計算機の発展により大量のデータを高速に処理可能になったことによって，統計や機械学習の分野で研究された学習手法が，扱うべきデータ量の多さのために今まで適用が困難であった一般物体認識へ適用できるようになってきた．人手によるルールやモデル構築から，統計的機械学習への移行は，人工知能や自然言語処理の研究でも見られる流れであり，大量データの高速処理が可能となったために可能となったアプローチであるといえる．それにともなっ

図 2 (a) 単語付きの Corel 画像の例．(b) Translation model による画像領域へのアノテーションの例．図は文献 40) より引用

Fig. 2 (a) An example of Corel images and their associated keywords. (b) An example result of image annotation by the translation model. The figure of the annotation result is cited from Ref. 40).

て，大量のデータから有用な知識を発見するデータマイニングという研究分野も確立された．一般物体認識は，単に画像認識やコンピュータビジョンの一研究分野というだけではなく，データマイニング，機械学習の応用分野にもなってきている．そのため，近年，一般画像認識の研究成果が CVPR，ICCV，ECCV などのコンピュータビジョン系の会議，前章で述べたマルチメディア系の会議に加えて，NIPS（Neural Information Processing Systems）や ICML（International Conference on Machine Learning）などの機械学習系の会議でも発表されるようになってきている．

さて，次に本章では，近年の一般物体認識ブームのきっかけとなった統計的機械学習手法を用いた研究について，代表的な方法を 2 つ述べる．

(1) **領域に基づく方法．**

(2) **局所パターンに基づく方法．**

(1) は，画像に自動的にキーワード付けを行うアノテーションのための方法で，かつての領域分割＋ラベリング規則の物体認識の方法に統計的学習手法を発展させたものであると同時に，データベースやマルチメディアのコミュニティで行われてきた画像検索の延長線上にある研究でもあるともいえる．画像 1 枚 1 枚をクラス分類するのではなく，データベース中の大量の画像に複数のふさわしいキーワードを付けるために提案された手法である．

(2) は純粋に従来の物体認識から発展した手法であり，それまでの物体認識の問題点を解決した新しい方法である．学習画像中の学習対象の切り出しが不要で，オクルージョンの問題にも対処可能である．この方法によって，一般物体認識の研究が 1 つの山を越えたといってもよいほど有望な手法である．

3.1 領域に基づく方法

領域に基づく方法で最も有名な方法が Barnard らによる word-image-translation model [32),33),40)]（以後，単に translation model と呼ぶ）である．彼らは，あらかじめ画像全体に対して数個のキーワードが付けられている Corel 画像データベースを用いて，領域分割された画像の領域への自動アノテーションを行った．具体的には，Blobworld [31)] もしくは Normalized Cuts [41)] を用いて領域分割し，画像と単語の対応のみで領域と単語の対応付けがされていない学習データを用いて，領域分割された各画像領域と単語の対応付けを統計的に推定する手法を提案した（**図 2**）．文単位で対応のとれている 2 カ国語で書かれた大量の文書（対訳コーパス）だけから，事前に辞書も文法の知識もなしに確率モデルによって辞書と文法を自動的に学習し機械翻訳を行う統計的機械翻訳[42)] の手法を画像に応用して，画像領域と単語の自動対応付けを実現している．

領域分割によって画像から切り出したすべての領域を一方の言語で書かれた文，画像に付けられた複数の単語をもう一方の言語で書かれた文と見なし，単語が付与された画像を大量に用意することによって，確率モデル（translation model）を学習し，画像の部分領域へのアノテーションを実現した．

Translation model における確率モデルの定式化の方法は数通り提案されているが，ここでは文献 43) で説明されている最も簡単なモデルを紹介する．確率モデルは，以下の式で表される領域特徴量 r に関する単語 w の条件付き確率分布によって表現される．

$$P(w|r) \propto P(w,r) = \sum_c P(w|c)P(r|c)P(c)$$

w は単語，r は領域の特徴量で，c はガウス分布のインデックスである．$P(r|c)$ は単一のガウス分布によって表現され，$P(c)$ はそれぞれのガウス分布の重みである．つまり領域の特徴量の分布はガウス混合分布（Gaussian Mixture Model，GMM）によって表現され，パラメータは EM アルゴリズムによって求められる．一方，$P(w|c)$ は各単一ガウス分布 c に対応す

る単語の確率で，学習データの各領域のガウス分布 c への帰属割合を重みとして，領域に対応する単語の頻度から計算する．なお，学習データは画像全体に単語が付与されていることを前提とするので，各領域にはその画像に付与された単語すべてが付与されていると仮定して学習する．

このように translation model では，単語ごとに GMM を求めるのではなく，すべての領域特徴の分布からグローバルな GMM を求め，単語付き画像データから求めた GMM 中の単一ガウス分布 c が単語 w に対応する確率 $P(w|c)$ を用いて，グローバルな GMM から単語ごとのローカルな GMM を生成して，確率モデルとしていると考えることができる．

一度確率モデルが求まると，単語が付与されていない画像から抽出された領域の特徴量 r に対して単語 w が対応する確率を計算することができる．なお，領域の特徴量ベクトルは色，形状，テクスチャなど，ごく一般的な特徴量が用いられている．

文献 33) では，領域の特徴量 r をベクトル量子化によって連続値から離散値に変換し，$P(r|c)$ を $P(w|c)$ と同様に学習データから出現頻度をカウントすることによって求める discrete translation model も提案されている．これは，後ほど 4.1 節で詳しく説明する文書分類のための確率的トピック抽出の手法 probabilistic Latent Semantic Analysis（pLSA）[44] と完全に等価な確率モデルになっている．実際，translation model の最初の論文[32] には，pLSA の提案者の Hofmann らが pLSA を提案する前に書いた論文[45] に影響を受けたと書かれているので，ルーツは同じであるといえる．

実際には，文献 32) が発表される以前に同様の考え方が日本人研究者によって発表されていた．森らは百科辞典中の画像と説明文から画像の部分領域と単語の対応を自動的に学習する手法を提案した[46),47]．この研究は文献 33) で引用されることによって，確率モデルの学習による画像への自動アノテーションのさきがけ的研究として世界に知られることとなった．方法としては，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ベクトル量子化の方法によっ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あらかじめ求めておく．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．この手法は一般に co-occurrence model と呼ばれている．

同じ手法を Web から収集したテキストと画像に対して行った研究[48] もある．ほかに類似研究として，Fung ら[49] も同様にクラス既知の学習画像をブロック分割，ベクトル量子化し，画像を量子化された各ブロックの組合せによって表現して，各クラスの平均的な量子化されたブロックの組合せを求めた．こ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 picture words を同様にして求めて，最も picture words が類似しているクラスに分類した．

以上述べた研究は，1980 年代にさかんに行われた領域分割と人手によるルールを用いた認識とは，学習データからシステムが自動的に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，単語付きの大量の画像データから学習した確率モデルによってシステムが自動的に推定する．

Translation model は ICCV2001 でオーラルペーパー[32] として発表されて，さらに ECCV2003 で best paper award in cognitive vision [40] を獲得して，最初は注目されたものの，物体認識のコミュニティにおいては，現在はあまり注目されなくなってしまっている．これは初期の領域分割の結果にその後の処理が依存してしまうことが大きな理由で，領域分割が容易な比較的単純なシーン認識には有効であるが，領域分割が困難な画像中の物体の認識には有効でないという問題点がある．認識結果を元に領域を統合する試み[43] も提案されているが，現在のところは次に述べる領域分割を行わない局所特徴量による方法の方が有効であると考えられている．しかしながら，translation model はマルチメディアのコミュニティでは依然として人気がある．なぜなら，領域に直接ラベルが付くことは結果が目に見えて分かりやすいからである．実際，CVPR では translation model のような領域を用いた研究は発表されていないが，ACM Multimedia や情報検索の国際会議の ACM SIGIR では translation model を改良や応用した研究がいくつか発表されている[50)〜53)]．

3.2 局所特徴量による方法

領域分割による方法では，オクルージョンがある場合や，形状が複雑で領域分割がうまくいかない場合には，対処することが難しい．そこで，Schmid らは，領域分割を行わずに，画像中の局所的な特徴の組合せによって，画像の照合を行う方法を提案した[56]．具体的には，最初に Harris 特徴点オペレータ[57] によって，画像中から 100 点程度の特徴点を選び出し，次に，各点の画素値や微分値などを特徴ベクトルとし，それらの集合によって 1 枚の画像を特徴付けることとした．照合は，未知の画像に対して，同様に特徴ベクトルの

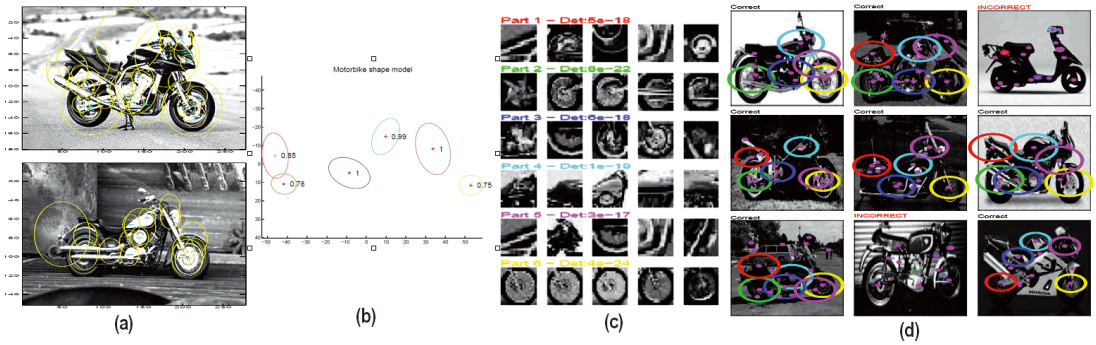

図 3 (a) バイク画像に対する Kadir-Brady detector[54] による検出結果．円の大きさは特徴点のスケールの大きさを示す．(b) 学習された部分の相対位置関係モデル．この例では 6 つの局所特徴から「バイク」モデルを構築している．(c) 局所パターン．この図は 5 枚の認識対象画像から検出された部分画像．(d) 認識結果．図は文献 55) より引用

Fig. 3 (a) Results of keypoint detection by Kadir-Brady detector[54] for "bike" images. The size of a circle corresponds to scale of the keypoint. (b) Trained spatial relation model. In this example, the "bike" model consists of six local parts. (c) Local patterns that are extracted from five "bike" images automatically. (d) Recognition results. The above figures are cited from Ref. 55).

集合を求めて，モデル画像（または学習画像）の特徴ベクトルの中から，それぞれ近い特徴ベクトルを探して，ある程度類似しているモデル画像に対して投票を行う．この際，特徴点間の相対的位置関係を考慮することによって，無駄な投票を防ぐことを行い，最終的に最も多くの投票を集めたモデル画像にマッチしたと見なす．特徴点自体は 1 画素の点にすぎないが，特徴点付近の画素から得られる高次元特徴ベクトルで特徴点を特徴付けることによって，点の集合による画像の表現が可能となった．この Schmid らの研究が，特徴点抽出アルゴリズムを用いた自動的な局所領域の切り出しによる物体認識の最初の研究であり，従来は 3 次元復元のための対応点抽出に使われていた特徴点抽出アルゴリズムが物体認識にも使えることを示したという点で重要な研究であるといえる．Lowe も同様の考えによって，独自に提案した特徴点抽出とその記述法を合わせた手法である SIFT（Scale Invariant Feature Transform）を用いて，オクルージョンのあるシーンにおける物体認識を実現している[58]．ただし，これらの研究は同一対象を探す identification の物体認識である．SIFT は当初，identification のために提案されたが，スケールや回転に不変な性質のために，現在では classification においてもその有効性が示されている[59]．詳しくは次章で説明する．

一方，Burl[60),61)] らは，局所領域の特徴とその位置関係を確率モデルで表現する constellation model（星座モデル）を提案した．この研究では classification の物体認識を実現していたが，学習画像の局所領域は人手であらかじめ指定しておく必要があった．それに対して，その発展研究の Weber らの研究[62),63)] では，constellation model に Schmid らの提案した特徴点抽出を局所領域の抽出手法として用いる方法[56)] を導入した．まず多数（300 枚程度）の正例，負例の両方の学習画像から Förstner 特徴点オペレータ[64)] を用いて局所パターンを抽出し，それらをクラスタリングすることによって対象に特徴的な局所パターンを選び出す．次に，局所パターンの見え方（appearance）と局所領域の位置関係を確率モデルで表現し，人間の顔や自動車の認識を学習によって実現している．利用する画像特徴は特徴点周りの局所パターンのみで，11×11 の濃淡パターンを用いている．なお，モデル名の constellation（星座）は，確率モデルが局所パターンとその位置関係をモデル化することに由来している．

これらの研究を発展させて，より多くの種類に対応可能として一般化したのが，CVPR 2003 で best paper になった Fergus らの研究[55)] である．オペレータは，特徴点周辺のパターンのスケール情報も出力される Kadir-Brady detector[54] が用いられた．これによって，ある程度の幅でのスケール変化にも対応可能となった．**図 3** に文献 55) での「バイク」の認識の例を示す．図 3 (d) を見ると，バイクの向きとスケールがほぼ揃えられているものの，様々なタイプのバイクを認識することができて，クラス内の変動に柔軟に対応できていることが分かる．

文献 55) における P 個の局所パターンからなる constellation model の確率モデルは，局所特徴のア

ピアランス D，それらの相対位置 X，局所特徴の配置のスケール S の同時確率として以下の式で表現される．

$$P(D,X,S) = \sum_{h\in H} P(D,X,S,h)$$
$$= \underbrace{P(D|h)}_{Apperance}\ \underbrace{P(X|S,h)}_{Shape}\ \underbrace{P(S|h)}_{Scale}\ \underbrace{P(h)}_{Combination}$$

D は P 個の局所特徴のアピアランス，X は P 個の局所特徴の相対位置座標，S は局所特徴の配置のスケール，h は画像から抽出された N 個の局所特徴を P 個のモデルの局所パターンに割り当てるインデックス，をそれぞれ表す．H は h のすべての組合せを表し，オクルージョンでモデルに対応する局所特徴がない場合も含めて組合せの数のオーダは $O(N^P)$ となる．

$P(D|h)$ の項は，局所特徴のアピアランスの確率モデルで，位置，スケールとは独立であると仮定されている．各局所パターンのアピアランスは単一のガウス分布で表現され，それぞれ独立であると仮定されるので，$P(D|h)$ は P 個のガウス分布の積として表される．$P(X|S,h)$ の項は局所特徴の位置関係の確率モデルで，つまり，認識対象の形を表現していて，P 個の局所特徴の x，y 座標の組を 1 つの $2P$ 次元ガウス分布で表現する．$P(S|h)$ はスケールの確率モデルでこれも単一のガウス分布で表現される．

これらのパラメータは translation model 同様，EM アルゴリズムによって求めるが，組合せが $O(N^P)$ であり，学習に時間がかかるのが欠点である．文献 55) によると $P = 5$〜7，$N = 20$〜30，学習画像データ 400 枚で，1 つのクラスを学習するのに 24 時間から 36 時間もかかったそうである．

そこで，文献 65) では，基準点を固定することによって上記の式の h の組合せ数を減らしモデル構築を高速化して，さらに主に瓶の認識のために物体の輪郭の曲線を局所特徴として導入した．また，Fei-Fei ら[66] は，constellation model において，他のクラスの学習データによって事前に構築した学習モデルを利用して，1 枚から 5 枚というわずかな学習画像で新しいクラスの確率モデルを構築する方法を提案した．以上の研究では，すべて濃淡画像を対象に認識が行われている．Translation model ではカラーが領域の特徴量として大きなウェイトを占めていたのとは対照的である．

Part-based による方法は，Perona 以外のグループによっても行われていて，Leibe ら[67] による局所パッチと一般化ハフ変換による方法や，Crandall ら[68] による constellation model によく似た k-fan と呼ばれる局所アピアランスと位置関係の確率モデルによる方法がよく知られている．

4. 一般物体認識における最新トピック

本章では，3 章で述べた局所特徴量を利用した一般物体認識をさらに発展させた最近の研究トピックについて解説する．なお，本章で解説する最新トピックを含めた近年の統計的手法に基づく一般物体認識の研究をまとめた論文集 Toward Category-level Object Recognition[69] が Springer の LNCS として 2006 年 12 月に出版された．本論文で触れていない研究も含んだ最新の多くの一般物体認識に関する研究論文，さらには物体認識研究の歴史，データセットに関する問題などの解説論文も含めて全 31 編の論文が収録されている．一般物体認識に興味のある方は，ぜひこちらを一読することをお勧めする．他の一般物体認識の最近のサーベイとしては，Pinz による文献 70) や Bosch らによる文献 71) がある．Datta らによる文献 72) は画像検索のサーベイであるが，画像アノテーションや画像分類などの一般物体認識に関連する内容も含まれている．

4.1 Bag-of-keypoints

Constellation model では，5〜8 個程度のごく少数の局所特徴量とそれらの位置関係によって物体を表現していたが，近年，位置情報をいっさい使わず，数百個のオーダの多数の局所特徴量の集合のみによって認識物体を表現する Bag-of-keypoints[73] と呼ばれる新しい手法が提案されている．

Bag-of-keypoints[73] は，統計的言語処理における bag-of-words model[74] のアナロジで，bag-of-words で語順を無視して文章を単語の集合と考えるのと同様に，bag-of-keypoints では位置を無視して画像を局所特徴（keypoints）の集合として考える．実際の処理においては，局所特徴の特徴ベクトルをベクトル量子化することによって，keypoint を word として扱えるようにする．このベクトル量子化された特徴を visual word もしくは visual alphabet と呼ぶこともある．つまり，bag-of-keypoints では，画像の特徴量は，画像から抽出した 100〜1,000 個程度の visual word の出現頻度のヒストグラムによって表現される．なお，この bag-of-keypoints と constellation model をまとめて，部分的な特徴を用いる方法ということで part-based approach と呼ぶ．

画像の局所パターンによるヒストグラム表現は，かつて電総研で研究されていた高次自己相関特徴量[75),76] と基本的には同じアイデアであるが，参照パターンが出現分布に応じて自動的に選ばれる点と，ヒストグラムの次元数が通常 100〜1,000 と大きい点が異なっている．

Bag-of-keypoints は統計的言語処理の bag-of-words

のアナロジであると最初に述べたが，画像の表現として bag-of-keypoints を用いることによって，統計的言語処理の分野で提案された確率的な手法が応用可能となる．たとえば，文書分類のための確率的トピック抽出の手法として提案された probabilistic Latent Semantic Analysis（pLSA）[44),77),78)]，Latent Dirichlet Allocation（LDA）[79),80)] などが一般物体認識に応用されている．

Latent Semantic Analysis（LSA）[81)] は bag-of-words によって表現された多数の文書集合から特異値分解によって文書集合の代表的なトピックを抽出する手法であるが，これを確率的な意味で再構築した手法が pLSA である．pLSA では潜在トピックを表す確率変数を導入し，単語と文書の出現確率を潜在トピックの混合分布としてモデル化する[44)]．LDA は pLSA に改良を加えた手法で，pLSA が確率モデルの表現，パラメータ推定に多項分布，EM アルゴリズムを用いるのに対して，LDA ではディリクレ分布，変分ベイズ学習[82)] をそれぞれ用いる点が異なっている[79)]．

3.1 節で述べたように pLSA と translation model はそのルーツが同じである等価なモデルであり，LDA もすでに画像のアノテーションに応用されている[33),50)]．このように pLSA や LDA の画像認識への応用は bag-of-keypoints 登場以前にもすでに試みられていたが，画像認識でのその有効性が広く知られるようになったのは，bag-of-keypoints に応用されるようになってからである．Translation model が統計的機械翻訳という自然言語処理の手法の応用であったことも含めて，一般物体認識の問題が言語の意味的処理の問題と共通する点が多いという点は大変興味深い．

Fei-Fei ら[80)] は，bag-of-keypoints 表現を用いて画像分類を行った．局所パターンを Lowe によって提案された SIFT 記述子（Scale Invariant Feature Transform descriptor）[58),83)] で表現し，13 クラスの学習画像 650 枚分の画像のすべての特徴量を k-means クラスタリングして，174 種類の code book（visual word）を作成した．画像はこの 174 種類の visual word の集合（bag）として表現される．確率的文書分類手法の LDA[80)] を用いて，13 種類のシーンを 64%の精度で分類した．従来は，形のある物体に対して part-based が主に試されていたが，物体認識に加えて，シーン認識にも有効であることが示された．特徴点オペレータを含む 4 つの方法で特徴点抽出を行っているが，山や海などの自然風景シーンはエッジやコーナが少なく特徴点オペレータによる特徴点抽出があまりうまくいかないようで，かつて Photobook[28)] などで行われたグリッド分割を用いた結果が最も良い結果になっている点も興味深い．なお，SIFT[58),83)] は，(1) 特徴点とそのスケールの検出，(2) 特徴点の 128 次元ベクトルによる記述の 2 つの処理を含んだアルゴリズムであるが，近年の一般物体認識では，(2) の特徴点の記述法のみを利用することが多く，Fei-Fei ら[80)] でも特徴点とそのスケール情報は別の方法で抽出しておいて，特徴点記述のみを SIFT で行った．SIFT の (2) のみの処理を用いて生成された特徴を本論文では以後「SIFT 記述子」と呼び，SIFT の (1)，(2) によって生成された特徴を「SIFT 特徴」と区別して呼ぶこととする．

Bag-of-keypoints では通常，局所特徴間の位置関係は考慮しないが，Fergus ら[77)] は平行移動とスケール変化に影響を受けないようにして位置情報を考慮するように pLSA[44)] を物体認識向けに改良した Translation and Scale Invariant pLSA（TSI-pLSA）を提案した．わずかではあるが認識率が向上した．

Visual word の考え方は，元々は identification の認識で提案された．Sivic らはビデオ映像から視点の異なる同一シーンを検索可能なシステム Video Google を提案した[84)]．SIFT 特徴[58)] をベクトル量子化し visual word を作成し，ビデオ中の各フレーム画像は多数の visual word を含んでいると考えた．そして，テキスト検索の手法を応用し高速な検索を実現した．

現在，一般画像認識においては，局所特徴量の記述法としては SIFT 記述子[58),83)] が最もよく用いられている．これは，その性能の高さに加えて，提案者の Lowe 自らによるものをはじめ，SIFT++[85)] などいくつかのソフトウェアが Web 上に公開されており，手軽に利用可能であるからである．物体認識における SIFT 記述子の性能に関しては，Mikolajczyk ら[59)] によって，局所特徴の記述子の中で SIFT 記述子が平均的に最も良い性能を示すことが報告されている．

Bag-of-keypoints 表現では，特徴点の表現は SIFT 記述子を用いるのが一般的であるが，特徴点の抽出方法には特徴点オペレータや，領域分割結果の領域の中心点，グリッド，ランダムなど様々な方法がある．Nowak ら[86)] は，様々な方法で特徴点を抽出し画像分類の結果を比較した結果，驚くべきことに，ランダムによる特徴点抽出が平均的に最も結果が良かったという興味深い報告を行っている．

Bag-of-keypoints 表現では，画像全体をそのまま特徴ベクトルとするため，認識対象物体のみでなく背景も含めて認識の手がかりとしていると考えられる．文献 87) では，画像認識コンテストの PASCAL Challenge[88)] において “やさしいデータセット” として提

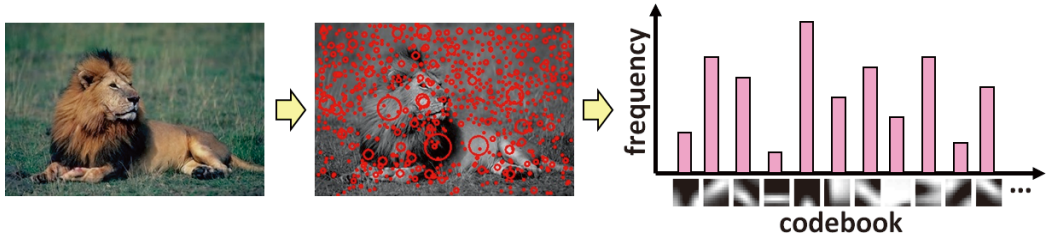

図 4 Bag-of-keypoints 表現の求め方．画像から SIFT によって特徴抽出し，コードブックに関するヒストグラムを作成する．ヒストグラムが画像の特徴量となる
Fig. 4 How to obtain bag-of-keypoints representation. Detect keypoints, extract SIFT vectors and build a histogram based on the pre-computed codebook. The histogram is regarded as a feature vector of the image.

供されている test1 画像セットの対象物体領域にマスクを掛けて，背景画像のみで 1/0 の画像分類を行ってみたところ，高い精度で分類が可能であったことが報告されている．

一般的には，Bag-of-keypoints による特徴表現は，学習画像に対する，(1) 特徴点抽出（100 個以上/画像），(2) SIFT 記述子ベクトルの計算，(3) 全学習画像の全 SIFT 記述子ベクトルの k-means によるクラスタリング（k は 100〜1,000 程度）による code book の作成，(4) code book に基づいて各画像について SIFT 記述子ベクトルのヒストグラムを作成，の手順によって作成することが可能である（**図 4**）．SIFT 記述子の計算以外は簡単な処理で，SIFT 記述子の計算も公開ソフトを利用することが可能なため，bag-of-keypoints はきわめて手軽な一般物体認識の画像表現であるといえる．

Bag-of-keypoints は，Jurie ら[89] による k-means によるクラスタリングの代わりの，オンラインクラスタリングと mean-shift[90] に基づいたコードブック作成法，Perronnin ら[91] による GMM および EM アルゴリズムによる確率的クラスタリングによるコードブック作成法，Weijer ら[92] による色情報の追加など，いくつかの改良が提案されている．

ICCV 2005 でのチュートリアル “Recognizing and Learning Object Categories” のホームページ[93] に part-based による一般物体認識の研究が Matlab のサンプルコード付きで詳しく解説されているので，詳しく知りたい人は参考にするとよい．

4.2 生成モデルから判別モデルへ：SVM の利用

Part-based アプローチが広まった当初は，constellation model や他の多くの part-based の研究が確率モデルによって表現される生成モデル（generative model）を認識モデルとして用いていたが，2006 年の CVPR で発表された 6 つの主な一般物体認識の研究[94)〜99)] のうち，constellation model の研究グループの Fei-Fei らの研究[96] 以外はすべてサポートベクタマシン（Support Vector Machine，SVM）に代表される判別モデル（discriminative model）を利用していた．Part-based は確率モデルのみでなく，工夫することによって SVM などにも適用でき，そちらの方がむしろ認識性能が良いということが分かってきた．これは 2006 年になってからの，新しい一般物体認識のトレンドになっている．

一般に，生成モデルに基づく方法では，あるカテゴリに属する画像の特徴量の分布モデルを学習データを用いて最尤推定で推定し，未知のデータに対して分布モデルを利用して事後確率を計算し，事後確率最大化（Maximum A Posteriori，MAP）推定によってどのカテゴリに属するか決定する．多くの場合は，モデル表現に混合分布を用いるために，そのパラメータを解析的に求めることが不可能で，反復によって近似解を求める手法である EM アルゴリズムを利用することが一般的である．一方，判別モデルに基づく方法では，学習データを用いて特徴ベクトルの空間を物体カテゴリに属する空間とそうでない空間に分割し，判別モデルを構築する．未知のデータは，分割した空間のどちらに属するかによって，物体カテゴリに属するかどうか判定される．SVM が判別モデルとしては一般に最も高い性能を示し，しかもその実装がフリーソフトウェア（SVMlight[100] や LIBSVM[101]）として容易に入手できるため，広く使われている．

SVM は高い性能を持ったクラス分類手法であるため，従来から様々な画像認識問題に応用されており，part-based アプローチにも近年，導入が試みられている．SVM を part-based アプローチに導入するには，画像間の類似度を計算するカーネル関数を定義することが必要であるが，bag-of-keypoints 表現は 1 つの画像全体を 1 つのベクトルで表現するので，同じ物体が含まれていても背景が異なればベクトル値は異なった

値となる．そこで，Grauman らは Pyramid Match Kernel[102] という 2 つの bag どうしの部分マッチングに基づいて類似度を計算するカーネル関数を提案し，bag-of-keypoints approach において SVM を用いた画像分類を行った．Lazebnik ら[95] は，Pyramid Match Kernel[102] に局所特徴の位置も考慮するように改良を加えた Spatial Matching を提案し，きわめて良い性能を示すことを示した．一方，Zhang ら[103] は，bag-of-keypoints を多次元ベクトルで表現するのではなく，多次元ベクトルとその重みの組の集合である signature で表現し，signature 間の距離を計算する方法である Earth Mover's Distance（EMD）[104] を SVM のカーネル関数として用いた．Signature 表現では，あらかじめ求めたコードブックに基づいて作成される bag-of-keypoints のヒストグラムとは異なり，ベクトル空間を画像ごとに k-means 法によって分割し，分割された空間の中心ベクトルとそれぞれの重みの組で，SIFT 特徴の分布を表現するため，表現がコンパクトになるが，EMD を求めるには線形計画問題を解く必要があるために距離計算に時間がかかるという問題点がある．

一方，constellation model に SVM を取り入れる研究[105),106] も行われている．文献 106) では，確率生成モデルから導出されるカーネルで，生成モデルを判別手法で利用するために一般的に利用される Fisher kernel[107] を用いて，constellation model を Fisher kernel 化し SVM を用いた分類を行った．実験では，従来の generative な方法[55] に対して性能向上が見られた．

Zhang ら[94] は，最近傍分類法（Nearest Neighbor）と SVM を組み合わせた SVM-KNN 法を提案した．SVM-KNN は簡単にいうと，まず K-NN 探索を行い，K 個がすべて同じラベルの対象ならそのクラスに分類し終了．そうでなければ，マルチクラス SVM を実行する．実際に一般物体認識における標準的な評価データ Caltech-101 に対して 101 種類の物体の分類を行い，今までのどの研究よりも良い結果を出している．ただし，特徴量が他の研究と異なり，独自の局所特徴抽出手法[108] を採用しているため，性能向上は SVM-KNN 法によるものだけでなく，独自の局所特徴抽出手法[108] によるものも少なくないと思われる．

このように，最近の一般物体認識の分類手法は，確率モデルによる生成手法から，判別手法に変化しつつある．

4.3 シーンの空間コンテキストの利用

Part-based の方法は基本的に単独の物体，単一のシーンを認識するのに用いられたが，実世界のシーンの画像中には複数の物体が含まれ，それぞれが何らかの関係を持って存在しているの普通である．たとえば，緑の草原の中に鉛直の棒状の物体があれば木である可能性が高いし，周りにビルがあれば電柱である可能性が高い．このように，物体単独の画像中でのアピアランスでは認識が困難で part-based では対処できない場合でも，画像中の他の部分が認識できれば，それとの関係から認識可能となる場合がある．こうした物体間の関係を利用した認識は空間コンテキスト（spatial context）を用いた認識と呼ばれている．空間コンテキストの利用はかつての知識ベース型システムでは当然のものとして利用されていた．たとえば，context-based recognition[16]，The Schema System[7] などは領域間の関係から認識を行っていた．ただし，ルールを人手で記述する必要があった．

最近，空間コンテキストを確率モデルによって表現し，学習によってモデルを構築する研究が行われるようになって，認識における空間コンテキストの利用に再び注目が集まっている．

Torralba ら[109] は，確率モデルをグラフ構造で表現するグラフィカルモデルを用いて，研究室シーンの画像に対して，desk，keyboard などの認識を行った．Sudderth ら[110] や Kumar ら[111] も似た方法で，part，object，scene の関係をグラフィカルモデルを用いて確率モデルとして表現し，路上シーンや研究室シーンの画像の認識を行った．

Hoiem ら[112] は，消失点を用いた簡単な 3 次元復元を行い，ベイジアンネットワーク[113] を用いて視点位置，地面，空，垂直領域，歩行者，自動車の関係をモデル化し，街中のシーンの画像に対して歩行者や自動車を認識した．この研究では，Marr 以降にさかんになった 3 次元復元を近年の統計的学習手法による一般物体認識に取り入れるという新しい興味深いアイデアを提案しており，2006 年の CVPR の best paper になっている．

これらの研究では，研究室シーンや路上シーンなど，1980 年代の研究で対象とされたシーンを対象としているが，かつてのような人手によるアドホックなルールに基づくのではなくて，確率変数の依存関係をグラフで表現するグラフィカルモデルを用いて，学習によって確率モデルを構築している．認識対象をグラフ中のノードとし，たとえば「自動車は道路の上にある」というような依存関係がある対象どうしのノードをエッジで結んで表現することによって，シーン中の依存関係を明確化し，各要素の確率モデルのみでなく，エッ

ジで結ばれた要素間の同時確率を求めることによって，シーン全体として最も事後確率が高い解釈を最終的に選ぶことにより認識を実現している．

以上述べたグラフ表現による確率モデルを用いた空間コンテキストの利用の研究はまだ始まったばかりで，対象シーン中に現れる可能性のある個々の物体がそれぞれ認識できる必要や，確率グラフの構造が学習でなく，人手であらかじめ与えているなどの理由によって，対象シーンが路上シーンや室内シーンなど狭い範囲に限定されているという問題点があり，今後の研究の発展が期待される．

ほかには，空間コンテキストではないが，一種のコンテキストの利用ともいえるシーン内の文字情報を利用した一般物体認識の研究[114] もある．たとえば，カメラならカメラメーカを表すロゴ，飛行機なら航空会社を表すロゴ，道路なら標識などが画像中に含まれており，これらを文字認識し，利用することで，物体の認識が容易になることが示されている．

4.4 シーン外のコンテキストの利用

前節で解説したコンテキストの利用は物体間の関係などのシーン内部の空間的なコンテキストであったが，画像の画素情報のみから画像内容を認識するという純粋な一般物体認識の限定を緩めることによって，様々な画像シーン外のコンテキストの利用が可能となる．

画像の撮影日時の時間的コンテキスト（temporal context）や，画像の撮影時のカメラ情報（絞り，シャッタースピード，焦点距離，フラッシュの有無など）のメタデータによって表現される撮影条件コンテキスト（imaging context），GPS で測定したグローバルな位置コンテキスト情報（spatial context），さらには一般の人が撮影した日常風景の画像であれば，その人の行動記録，日記，メール，ブログ，さらには閲覧した Web ページのログなど，様々な情報がすべて，画像認識のためのコンテキスト情報として利用可能である．

Boutell らは，同一の人が撮影した写真画像の撮影時間間隔を隠れマルコフモデルでモデル化して，時間的に連続して撮影された写真の自動分類を行った[115),116)]．また，同じ著者らは，デジタルカメラによって撮影された画像の JPEG ファイル中に Exif 規格に基づいて記録されているカメラ情報のメタデータを利用して，ベイジアンネットワークによってシーン分類することも行った[117)]．

5. 評価の方法とグランドツルースの作成

本章では，一般物体認識の結果の評価や，学習データセット構築に関する話題について触れる．

5.1 評価データセット

以上述べたように研究がさかんになると，各手法が比較できるように，統一した評価が重要になってくる．統一した評価を行うためには標準的な評価データセットが必要であるが，一般物体認識については，かつては，キーワードが 6 万枚の画像に対して付与されている Corel 社の Corel Image Gallery がデファクトスタンダードであった．実際，translation model[33)] をはじめとして多くの研究で評価に Corel 画像が用いられていた．しかしながら，Corel 社が Corel 画像の販売を数年前にとりやめてしまったために，現在は入手不可能であるという問題点と，元々画像認識のためのデータセットとして作られたわけではないので，対象によって認識の難易度の差が大きすぎるという問題点があり，現在ではあまり使われなくなっている．

そこで 2005 年以降は，Corel 画像に代わって，カルフォルニア工科大学の Caltech-101[118),119)] が評価画像データのデファクトスタンダードとなっている．元々は文献 118) で用いた実験データであるが，Web 上[119)] で公開されているために，誰でも入手可能である☆．この Caltech-101 画像セットは，その名のとおり 101 種類の画像からなり，主に Google Image Search を用いて人手で集めた 9,144 枚の画像から構成される．クラスごとに枚数が異なり，31 枚から 800 枚までとばらつきがある．Airplane，bike など人間の正面顔画像 face を含めて様々な物体の画像が含まれている．どれも物体画像で，風景画像は含まれていないため，物体の認識用のデータセットであるといえる．文献 55)，66) などで元々実験データとして使われていた face，airplane，motor bike はそれぞれ 870 枚，800 枚，798 枚と突出して多いが，それ以外はおおむね 50 枚前後である．図 3 (d) のバイク画像も Caltech-101 の一部であるが，このバイク画像のように，多くの Caltech-101 画像では物体の向きと大きさがほぼ揃えられているという点も特徴である[87)]．

2006 年の時点で Caltech-101 を使った画像分類において，最も良い結果を出しているのが UC Berkeley のグループの認識率 66.23%[94)] である☆☆．これは各クラスごとにランダムに 30 枚の学習画像を選び出して，残りをテスト画像とし，reject なしの 101 クラス分類を 10 回行った結果の平均値である．**表 1** に 2006

☆ 文献 118) での実験結果には zebra が含まれているが，公開されているデータセットにはなぜか zebra の代わりに faces_easy（認識が容易な顔）が faces とは別のクラスとして含まれている．

☆☆ 2007 年には 81.3%が報告されている[120)]．日本国内では 2007 年に 58.69%を出した結果が報告されている[121)]．

表 1 Caltech-101 の平均分類精度
Table 1 Reported classification rates on Caltech-101 dataset.

順位	グループ	発表学会	文献 no.	結果（%）
1	UCB	CVPR'06	94)	**66.23**
2	INRIA	CVPR'06	95)	**64.6**
3	UIUC	CVPR'06	96)	**63**
4	MIT	ICCV'05	102)	**58**
5	UBC	CVPR'06	98)	**56**
6	MIT	CVPR'06	99)	**51.2**
参考	Caltech	PAMI'06	122)	**17.7**

年の CVPR までに発表された上位 6 位までの結果を示す．このうち 3 位の文献 96) を除いては SVM を用いた判別手法による認識である．ちなみに，元祖の constellation model による Caltech-101 の分類結果は 17.7% [122] であり，ここの 2，3 年のレベルアップは目覚ましい．

ただし，文献 87) で指摘されているように，Caltech-101 は物体の向きと大きさがほぼ揃えられていてクラス内変動が小さく認識しやすい画像を集めた評価セットであるという批判が以前からあった．それに応える形で，2007 年になって画像枚数 30,607 枚の Caltech-256 [123] が公開された．Caltech-256 はカテゴリが 256 種類に増えただけでなく，Caltech-101 作成時に行った向きや大きさを統一するような操作を行わなかったため，より「一般的な」データセットになっているという特徴がある．このことは，Caltech-256 の作成グループが自ら Caltech-256 について分析し検証している [124]．文献 124) によると，Spatial Pyramid Kernel [95] を用いた方法で 256 種類分類のときは分類精度 34.1%で，256 種類からランダムで選んだ 100 種のみを使ったときの分類精度は平均 45%程度と，Caltech-101 に対する結果 64.6%に比べて大幅に低下している☆．

このように，一般物体認識は，認識対象のクラスの選び方，学習画像，評価画像の選び方によって認識結果が大きく変わるという問題がある．たとえば，かつてはシーン認識において sunset がよく用いられていた．夕暮れの画像は画像全体が赤いきわめて特徴的な画像であり，シーン分類が比較的用意であるので，分類クラス中に sunset を入れておくと全体の認識率の数値を上げることができたからである．こうした問題に対処するには，評価に統一した評価セットを用いることが重要で，その意味ではデータの良し悪しは別として Caltech-101 [118] は一般物体認識研究の統一的評価を可能とし，研究全体のレベルを向上させるのにおおいに貢献しているといえる．なお，こうした一般物体認識の研究において評価によく用いられる認識クラスを調査した研究 [125] もある．

5.2 ベンチマークワークショップ

Caltech-101 による統一的評価だけでなく，一般物体認識の手法を競うベンチマークワークショップというのも開催されている．これは，主催者の提供する共通の学習画像データとテストデータを用いて，共通の課題を処理し，結果を競うというものである．結果は良い方が望ましいが，厳密な意味でのコンテストではないので 1 位になっても表彰されることはなく，後日開かれるワークショップで参加者同士が自分の手法を発表してお互いに情報を交換し合うことで，コミュニティ全体の技術を向上させていくことが目的である．PASCAL Challenge [88]，TRECVID [126),127)]，ImageCLEF [128] が一般物体認識に関係した，誰でも参加可能なオープンなベンチマークワークショップとしてあげられる．それぞれ，Web 上でワークショップの予稿がすべて公開されており，詳しい情報を得ることが可能である．

PASCAL Challenge Visual Object Class [88] は，ヨーロッパのパターン認識および機械学習コミュニティの PASCAL（Pattern Analysis, Statistical Modelling and Computational Learning）によって主催されているコンテストの中の一般物体認識部門で，与えられた学習画像を用いて与えられたテスト画像から 10 種類の物体（bicycle, bus, car, cat, cow, dog, horse, motorbike, person, sheep）を認識する．課題は画像に含まれているかだけを判別する classification 課題と，画像のどこに含まれているかも detection 認識する課題の 2 つがある．Part-based の研究を行っているヨーロッパの主な研究者の多くはこれに参加しているようである．Caltech-101 はどちらかというと認識しやすい画像のみを集めているが，PASCAL Challenge の提供する画像は，一般のスナップ写真に近いもので，オクルージョンのある画像も含まれている．提供画像は全部で 2,800 枚程度で，枚数はあまり多くはない．

PASCAL Challenge の 2006 年の結果は classification 課題で最高 9 割以上，detection 課題で最高 4 割程度となっている．ただし評価方法は，Caltech-101 の標準的な評価方法と異なり，各クラスで 1/0 の 2 クラス分類を行った平均なので，Caltech-101 の結果とは比較不可能である．

☆ 2007 年には 256 種類の分類精度として 45.3%が報告されている [120]．

TRECVID [126),127)] は，アメリカの国立の技術標準化機関 NIST（National Institute of Standards and Technology）の研究部門が行うテキスト検索ワークショップ TREC（Text REtrieval Contest）から派生したビデオ映像検索ワークショップである．アメリカのニュース番組 CNN や NBC，中国語およびアラビア語のニュース番組など合計約 160 時間分の実際のニュース映像から決められた 39 の物体もしくはシーンを含むショットを選び出す高次特徴抽出課題（high-level feature extraction task）が一般物体認識に近い課題である．認識物体，シーンは映像検索を意識したもので，たとえば，explosion というシーンと car という物体の認識結果を組み合わせることによって，car explosion シーンの検索が可能となる．ほかにもショット分割課題，検索課題がある．高次特徴抽出課題では，主催者から提供される映像の分割単位であるショットを対象に対象物体，シーンを含む候補のショットを最大 2,000 まで解答する．ただし，静止画像ではなく映像が対象なので，音声，音声を自動音声認識した音声認識テキスト（ただし，中国語，アラビア語の映像の場合は，音声認識後，英語に翻訳したテキスト），ニュース映像中の字幕を文字認識した字幕文字認識テキスト，動き情報などが静止画に加えて主催者より提供される．そのため，対象によっては，画像認識よりもテキスト検索で十分対処可能な場合もある．また，逆に，各ショットの代表画像も与えられるため，映像であることを無視して，純粋に一般物体認識問題として取り組むことも可能である．ただし，テスト映像のショットは全部で 14 万以上もあり，時間のかかる認識手法は困難である．2006 年度以降は 14 万ショット以上の大規模映像データに対して 39 種類すべてについて認識を行わないとならないため，参加のための敷居が高いという問題がある．

認識結果は，2006 年の 39 課題のうち実際に評価された 20 課題（sports，weather，office，meeting，desert，mountain，waterscape，corporate leader，police，military personnel，animal，computer tv screen，US flag，airplane，car，truck，people marching，explosion fire，maps，charts）の場合，20 種類平均で，適合率の最高が 0.192，平均が 0.078 であった．TRECVID は映像が対象のため，参加者の多くは映像処理の研究者であるが，わずかに画像認識の研究者も参加している．2006 年は，著名な画像認識研究グループの参加チームとしては，Caltech-101 の分類において 2006 年の時点で最高の結果[94)] を出している UC Berkeley の Malik のグループ，Video Google [84)] の研究などで知られる Oxford の Zisserman のグループが参加していた．

UC Berkeley のグループは，Caltech-101 に対して用いた手法[94)] を TRECVID に持ち込んだ．音声認識テキストや動き情報などの映像特有の情報はいっさい利用せずに，TRECVID を純粋に一般物体認識の問題として取り組んだ場合の実験結果として，参加者全体の平均 0.078 を大きく上回る 0.110 の適合率を実現した[129)]．また，Oxford のグループは，画像特徴は Bag-of-keypoints 表現，分類法は Spatial Pyramid Match Kernel [95)] を用いた SVM を利用し，画像特徴のみで 0.093 の適合率を実現した[130)]．このことは最新の一般物体認識の手法は映像検索にも適用できることを示している．

なお，TRECVID 参加者は著作権に関する誓約書を書くことによって研究目的に限った映像の使用が自由となり，著作権問題がクリアされている．一方，Caltech-101/256 は Web から収集した画像が多く含まれており，しかも引用元の情報である画像の URL の情報が添付されていないなど，著作権の問題に関してはまったく考慮されていないという問題を含んでいる．

ImageCLEF [128)] は，多言語情報検索のワークショップ CLEF の画像検索の部門で，21 種類の物体を含む 1,000 枚のデータベース画像に対して画像分類を行う課題がある．2006 年度の認識結果は 2 割程度であまり高くはない．こちらも PASCAL Challenge 同様ヨーロッパで行われており，主に情報検索の研究者の参加が中心になっている．

5.3 人手による学習データ作成

次に，学習や評価に必要なグランドツルースデータの作成の問題について触れる．種類が少ない場合は研究者が独自に構築することができたが，大規模なグランドツルースデータセットを構築するためには研究者が共同で構築することが不可欠である．

現在は，Caltech-101/256 も PASCAL Challenge も TRECVID もすべて人手によって学習データおよび評価データが作成されている．今後，認識対象が千種類，1 万種類と増えるにつれて，学習データの作成が困難となってくることが予想される．評価はサンプリングによって行うことも可能であるが，学習データは一般に正解データ（グランドツルース，ground-truth）である必要があり，人手によって労力をかけて集めることが必要である．

Caltech-101/256 は画像枚数がそれぞれ 9,000 枚，30,000 枚程度なので Caltech のグループが独自に構築したが，画像データがさらに多くなると単独グルー

プが構築することは困難である．TRECVID では参加者が共同で，学習用映像から切り出された約 4 万枚の画像に対して 42 種類の物体/シーンのアノテーションを行い，グランドツルースの作成を行っている[131]．また，TRECVID のニュース映像データについて，人手によって 1,000 種類のグランドツルースを作ろうとしている IBM，CMU，U Colombia を中心とした LSCOM（Large-Scale Concept Ontology for Multimedia）[132] というプロジェクトもある．1,000 種類にもなると対象コンセプトを選ぶのも簡単でなく，言語の階層構造であるオントロジを考慮して利用価値の高い 1,000 種類のコンセプトの定義を目指している．

ほかに，大まかに領域分割された画像のそれぞれの領域にアノテーションしたグランドツルースデータを構築する LabelMe プロジェクト[133),134)] というものもある．こうしたデータは，画像全体にアノテーションされたものより構築に手間がかかり，画像中からの物体の位置の検出まで含めた認識のための研究データとして利用価値が高い．

面白い試みとして，グランドツルースデータ作成時に必要な画像へのアノテーションの作業自体をオンラインゲームにしてしまって，ネットワーク上の多くの人々の力を使って画像へのアノテーションを行う試みがある[135]．ESP game[136] は，CMU の学生の Ahn らが作った画像アノテーションのオンラインゲームサイトで，ユーザに提示した画像に対してその画像を表す単語を入力させることをゲームとしている．ゲームが人気になってユーザが増えれば，Web 上の多くの画像を簡単にラベル付けすることが可能で，現在，すでに 1,000 万枚以上の画像に連想されるキーワードが付けられているそうである．その一部の 30,000 枚については，現在 Web 上☆で公開されている．1 枚の画像に対して複数のプレーヤにアノテーションさせるので，正しい連想キーワードを多数決で決めることができ，それによってアノテーションの精度は実用レベルになっている[135]．このアイデアは Google に技術移転され，将来的に Google Image Search の精度向上に役立てることを想定して，Google Labeler[137] としてサービスされている．

なお，2006 年には ESP game をさらに発展させて，ESP game によって収集されたキーワードが画像のどの部分に対応するかをゲームプレーヤに示させる Peekaboom[138),139)] を提案している．これによって，LabelMe プロジェクト[133),134)] のように，画像中の対応する部分にアノテーションされたデータがオンラインゲームによって収集可能となっている．

5.4 自動による学習データ作成

一方，一般画像認識のための自動知識獲得の研究もある．知識源は World Wide Web である．

柳井[140),141)] は Web から分類クラスを表すキーワードを用いて画像を収集し，その Web から収集した画像を一般の画像分類のための学習画像として用いることを提案した．近年，Web からの知識獲得（Web マイニング）の研究がさかんに行われているが，文献 140)，141) では，それと同様に，Web 上の画像がテキストによる HTML ファイルからリンクされていることを利用して，実世界画像とその意味内容との対応の知識（ここでは画像知識と呼ぶ）を Web から自動的に獲得した．そして，その知識を実世界画像分類や自動キーワード付与などに応用することを提案し，こうした Web からの画像知識の獲得を「Web 画像マイニング」と呼んでいる．方法は，まず分類クラスに対応するキーワードに関連する画像を大量に Web から収集し，未知画像を最近傍法（Nearest Neighbor）で分類する．画像特徴量および距離は画像検索の手法である Earth Mover's Distance（EMD）[104] および Integrated Region Matching（IRM）[142] を利用している．20 クラスで 4 割程度の分類率であるが，20 個のキーワード入力のみで人手の介入なしに 20 クラスの画像の分類が可能となっている．

Constellation model[55] の提案者 Fergus らは，Google Image Search の結果の画像を用いて認識モデルの学習[77),143)] を行った．この研究では，Google Image Search の出力から RANSAC[144] の手法を用いてキーワードに対応する画像のモデルを人手の介入なしに学習し，10 種類のキーワードに対して，15%の再現率の場合 58.9%の適合率で画像選択が可能となっている．

ほかにも，Web 上の画像を用いた物体認識の研究[145] は存在しており，今後同種の研究は増加していくことが予想される．特に近年は Yahoo 画像検索 API[146] や，Flickr API[147] などの画像を容易に Web から収集するための Web API サービスが提供されるようになっており，こうした研究を行うための環境が整ってきている．

一方，AnnoSearch[148] では，Web 上のオンラインフォト Web サイトの 240 万枚の画像とユーザによって付加されたキーワードを知識として，類似画像検索を用いた自動画像アノテーションを提案している．文献 140)，141) では，あらかじめ分類クラスを指定す

☆ http://www.captcha.net/esp-search.html

る必要があったが，文献 148) ではその必要はなく，どのような画像に対しても平均 3 割程度の精度で，自動アノテーションを実現している．

Web 上の知識は人手によって構築されたグランドツルースとは異なり，つねに誤った知識（ノイズ）が含まれている．たとえば，ライオン画像を Web から収集しても，収集した画像の適合率は良くても 7〜8 割程度にしかならない．そこで，こうしたノイズを含む Web 上のデータを利用するためには，ノイズの除去が重要である．Fergus ら[143] はモデル学習時に RANSAC[144] を用いた．一方，Angelova ら[149] は，classification の物体認識を行う場合に不要な学習画像，不適切な学習画像を取り除く方法を提案して，今後の課題で Web 画像に適用予定と述べている．柳井ら[150),151)] は EM アルゴリズムを応用した繰返し手法によって，モデル学習時にノイズの影響を少なくする方法を提案している．

Fergus らによる文献 77) では，精度の高い Web 画像を取得するために，変わった方法を採用している．Google Image Search から学習画像を取得する際に，検索結果の上位 5 位以内にノイズがほとんど含まれないという経験則を利用して，機械翻訳を用いてクラスに対応するキーワードを英語以外の 6 カ国語に翻訳し，7 カ国語で多言語画像検索を行い，それぞれの検索結果の上位 5 枚の合計 35 枚を学習画像とした．

Web 上のデータはノイズをつねに含むために，人手による学習データには正確さではかなわないものの，人手によるデータ収集はかならずデータ作成者の意図が反映されてしまうという問題があるのに対して，Web 上の画像（Web 画像）は様々な人が様々な目的で撮影した画像であり，実世界の一般的な画像の多様性をそのまま反映していると考えられる．Web から画像およびそれに付随するテキスト情報を自動収集することによって，真に “一般的な” データセットが構築できる可能性がある．また，それとは別の問題として，**「Web から一般物体認識のための知識を自動獲得できるのか？」**という問題自体も興味深い研究課題である．

6. 今後の展望

Part-based 手法の提案によって，新たな局面を迎えた一般物体認識ではあるが，実用的に一般の人々に用いられるようになるまでには，今後解決すべき問題は多く残っている．本章では，以下の 2 つの問題について触れることにする．

- **多種類化と認識クラスの決め方．**
- **クラス内変化への対応．**

6.1 多種類化と認識クラスの決め方

今後は多種類への対応はますます進んで，1，2 年以内には 1,000 種類の認識が行われるようになることが予想される．実際に，LSCOM（Large-Scale Concept Ontology for Multimedia）[132] プロジェクトでは，1,000 種類のコンセプトを定めて，ニュース映像へのアノテーションを行おうとしている．我々のグループでも現在，1,000 種類のカテゴリの画像を Web から各 1,000 枚ずつ収集することを行っているが，1,000 種類になるとカテゴリに対応する名詞を選ぶだけでも簡単ではないことが分かってきている．それは，1 つの物体を表す名詞は多数あり，その中に認識に向いている名詞とそうでないものがあるからである．

実世界には認識対象は辞書に出ている（具象）名詞の数ほどあって，人間は数万種類の対象を認識可能であるといわれる[1]．ところが，認識すべきクラスに対応する名詞の概念は互いに独立ではなく，instance-of 関係，part-of 関係，made-of 関係などで互いに階層的な構造を構成している．つまり，「乗用車」は上位概念の「乗り物」でもあり，下位概念の「セダン」や「トヨタヴィッツ」でもあるかもしれないというように，乗用車という物理的実体を表すには多くの名称が存在して「乗用車」という名称はその中の一名称でしかない．これは instance-of 関係であるが，part-of 関係を考えると「乗用車」は「タイヤ」でも「車体」でも「窓」でもあるともいえ，また，made-of 関係を考えると「乗用車」は「鉄板」や「ゴム」「ガラス」などであるともいえる．

このため，言語の階層構造をつねに考慮して，認識すべきクラスに対応する名詞を選ぶ必要がある．Rosch ら[152] は，言語の階層構造のレベルでの認識が人間にとって最も基本的な認識であるかを心理実験によって明らかにしていて，人がぱっと見たときにすぐに思いついたり，幼児が最初に覚えたりするような，基本認識レベル（basic-level category）の考え方を提案している．基本認識レベルでは同一名称の対象は多くの共通の性質を持っていて，特に，(a) 形状の類似性，(b) 運動，動作，操作の類似性を持っている．ということを述べている．基本認識レベルはつまり認識しやすいレベルということであり，この考え方は一般物体認識での認識クラスを決める際に参考になる考え方である．

こうした問題に対して，我々は，単語が表す「概念」がどの程度，視覚的特徴を持ち合わせているか，つまり概念の視覚性（visualness）を定量的に評価する方法を研究している[153),154)]．我々は言語階層中の視覚

性が高い概念から優先して認識を行うクラスとして採用するべきであると考えている．

Sivic ら[78] は bag-of-keypoints approach を用いて，大量の画像に対して文書分類手法の probabilistic Latent Semantic Analysis（pLSA）[44] を適用することによって，自動的に画像のクラスを抽出する concept discovery を提案している．あらかじめ分類クラスを決めて，それに対応する学習画像を人手で集める従来一般的な supervised な方法とは異なり，大量の画像から自動的にクラスを探し出す unsupervised な方法の試みで，認識すべきクラスを自動発見するという興味深いアイデアを提案している．

6.2 クラス内変化への対応

Caltech-101 は，クラス数は 101 もあるが，実は同一クラス内の画像はバイク画像のように同じような見た目の画像を意識的に集めていて，同一クラス内の変化はあまり大きくない．物体に対する視点の方向は様々な場合が考えられるが，写真として撮影される場合の視点は限られている．バイクの場合，真上や真下から見ることはほとんどなく，横もしくは斜め横から見ることが普通である．そのため，横，斜め前方，斜め後など典型的な視点方向に対応できれば，かなりの場合に関して認識が可能となると思われる．

こうした人間にとって典型的な見え方を canonical perspective [155] といい，文献 155) では被験者を用いた心理実験によって物体の典型的なビューを調べている．顔画像認識が正面顔のみを対象にして成功していることからも分かるように，典型的なビューが認識できれば特殊な場合は認識できなくても，実用上は問題なく，そのクラス内では “一般的な” 認識ができたと認めることができるであろう．

どの方向からのビューが世の中で典型的であるかを多くの対象について調べるのは容易ではないが，Web 画像を利用することである程度行うことができる．Web から大量の画像を収集して，各クラスの典型的なビューをクラスタリングなどによって自動的に探し出す研究は実際に行われている．Fergus ら[77] は，Google Image Search から単一のキーワードで収集した画像を bag-of-keypoints 表現をし，テキストのトピック分類手法である pLSA [44] を物体認識向けに改良した Translation and Scale Invariant pLSA（TSI-pLSA）を用いて，自動的にビューごとに分類することを行っている．

クラスによってはクラス内での変動が大きい場合があり，これも困難な問題である．特に人工物は，その機能によって名称を付けているので，同一名称であっても見た目がまったく異なる場合が多くある．たとえば，「椅子」を考えた場合，椅子には 1 本足の回転する椅子もあれば，4 本足の椅子，ソファーのような椅子，公園のベンチのような椅子もある．これらをすべて「椅子」として認識するのか，サブクラスを作って別々に認識するかは問題になる．これは概念の視覚性（visualness）の問題とも関係してくる．

ほかに，画像中に多数の物体が存在している場合のオクルージョンの問題や，画像中の物体自体が小さく，十分な特徴が得られない場合の問題がある．オクルージョンは局所特徴量を利用することである程度対処可能であるが，大部分隠れてしまった場合や，小さくて十分な特徴が得られない場合は，シーンや物体間の空間コンテキストの利用が必要であろう．学習を用いた統計的アプローチによるコンテキストの利用は研究が始まったばかりで，今後の発展が期待される．ただし，コンテキストには，物体どうしの物理的な関係（上にものが載っているという関係），関連物体の同時存在（e.g. 机の前に椅子がある），スケールの関係（e.g. 手のひらの上のクルマはミニカーだが，同じ大きさでも遠方にあれば普通の自動車）など考慮すべき関係の種類が多く，これらを統一的に統計モデルで扱うことができる枠組みを実現することは容易ではない．

7. おわりに

本論文では，一般物体認識の過去から最新の動向までをまとめ，一般物体認識の解決すべき課題について考察した．

現在は，物体やシーンの名称で認識を行っているが，究極的には “One image tells many things.” を実現できるような認識システムが望まれる．つまり，含まれる物体やシーンの認識をするだけでなく，人間が行う「想像」のように画像から予測される様々な可能性について，システムが理解して語ることが望ましい．こうしたことが実現できて，初めて「画像の意味的な認識・理解」が実現できたといえるのではないかと考える．そのためには，クラスと画像特徴の対応の知識だけでなく，コンテキストに関する知識を含めた様々な種類の，人間が無意識に用いている視覚情報に関する「常識」をシステムが獲得する必要がある．このためには，膨大な視覚情報，特に Web 上にある画像情報からのデータマイニングや知識発見によって「常識」を獲得することが実現のための鍵となるであろう．そうなると「画像認識」は，もはや画像認識やコンピュータビジョンの枠にはとどまらず，様々な知識を総動員する，まさに究極の人工知能の問題になるといえる[156]．

謝辞 本論文の草稿に対してコメントしてくださっ

た和歌山大学の和田俊和先生，国立情報学研究所の井上雅史先生，電気通信大学の堀田一弘先生，東京農工大学の堀田政二先生，名古屋大学の神谷保徳さんに感謝申し上げます.

参 考 文 献

1) Biederman, I.: Human image understanding: Recent research and a theory, *Computer Vision, Graphics and Image Processing*, Vol.32, No.1, pp.29–73 (1985).
2) Ullman, S.: *High-level Vision*, The MIT Press (1996).
3) 柳井啓司：一般物体認識の現状と今後，情報処理学会コンピュータビジョン・イメージメディア研究会報告 CVIM2005-155-17 (2006).
4) Clowes, M.B.: On Seeing things, *Artificial Intelligence*, Vol.2, No.1, pp.79–116 (1971).
5) Tenenbaum, J.M. and Barrow, H.G.: Experiments in Interpretation Guided Segmentation, *Artificial Intelligence*, Vol.8, pp.241–274 (1977).
6) Ohta, Y.: *Knowledge-Based Interpretation of Outdoor Natural Color Scenes*, Pitman Advanced Publishing Program, Boston (1985).
7) Draper, B., Collins, R., Brolio, J., Hanson, A. and Riseman, E.: The Schema System, *International Journal of Computer Vision*, Vol.3, No.2, pp.209–250 (1989).
8) Matsuyama, T. and Hwang, V.S.: *SIGMA: A knowledge-based aerial image understanding system*, Plenum Press, New York (1990).
9) Marr, D.: *Vision*, Freeman (1982). 乾，安藤（訳）：ビジョン，産業図書 (1985).
10) Batlle, J., Casals, A., Freixenet, J. and Marti, J.: A review on strategies for recognizing natural objects in colour images of outdoor scenes, *Image and Vision Computing*, Vol.18, No.6-7, pp.515–530 (2000).
11) Pope, A.R.: Model-Based Object Recognition: A Survey of Recent Research, Technical Report TR-94-04, University of British Columbia, Computer Science Department (1994).
12) Binford, T.: Visual Perception by Computer, *Proc. IEEE Conf. on Systems and Control* (1975).
13) Brooks, R.A.: Model-Based Three-Dimensional Interpretations of Two-Dimensional Image, *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol.5, No.2, pp.140–150 (1983).
14) Basri, R.: Recognition by Prototypes, *International Journal of Computer Vision*, Vol.10, No.2, pp.147–167 (1996).
15) Stark, L. and Bowyer, K.: Achieving Generalized Object Recognition through Reasoning about Association of Function to Structure, *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol.13, No.10, pp.1097–1104 (1991).
16) Strat, T.M. and Fischler, M.A.: Context-Based Vision: Recognizing Objects Using Information from Both 2-D and 3-D Imagery, *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol.13, No.10, pp.1050–1065 (1991).
17) 松山隆司，尾崎正治：LLVE：トップダウンセグメンテーションのための画像エキスパートシステム，情報処理学会論文誌，Vol.27, No.2, pp.191–204 (1986).
18) 長谷川純一，久保田浩明，鳥脇純一郎：サンプル図形提示方式による画像処理エキスパートシステム IMPRESS，電子情報通信学会論文誌 D, Vol.J70-D, No.11, pp.2147–2153 (1987).
19) Clement, V. and Thonnat, M.: A Knowledge-Based Approach to Integration of Image Processing Procedures, *Computer Vision, Graphics and Image Processing*, Vol.57, No.2, pp.166–184 (1993).
20) Swain, M.J. and Ballard, D.H.: Color Indexing, *International Journal of Computer Vision*, Vol.7, No.1, pp.11–32 (1991).
21) 村瀬 洋，Vinod, V.V.：ヒストグラム特徴を用いた高速物体探索法—アクティブ探索法，電子情報通信学会論文誌 D-II，Vol.J81-D-II, No.9, pp.2035–2042 (1998).
22) Kashino, K., Kurozumi, T. and Murase, H.: A Quick Search Method for Audio and Video Signals Based on Histogram Pruning, *IEEE Trans. Multimedia*, Vol.5, No.3, pp.348–357 (2003).
23) Schiele, B. and Crowley, J.L.: Recognition using Multidimensional Receptive Field Histograms, *Proc. European Conference on Computer Vision*, pp.610–619 (1996).
24) Turk, M. and Pentland, A.: Eigenfaces for Recognition, *Cognitive Neuroscience*, Vol.3, No.1, pp.71–96 (1991).
25) Murase, H. and Nayar, S.K.: Visual Learning and Recognition of 3-D Objects from Appearance, *International Journal of Computer Vision*, Vol.14, No.9, pp.5–24 (1995).
26) Gudivada, V.N. and Raghavan, V.V.: Content-Based Image Retrieval-Systems, *IEEE Comput.*, Vol.28, No.9, pp.18–22 (1995).
27) 串間和彦，赤間浩樹，紺谷精一，山室雅司：色や形状等の表層的特徴量に基づく画像内容検索記述，情報処理学会論文誌：データベース，Vol.40, No.SIG3 (TOD1), pp.171–184 (1999).
28) Minka, T.P. and Picard, R.W.: Vision Texture for Annotation, *ACM/Springer Journal of Multimedia Systems*, Vol.3, pp.3–14 (1995).

29) Belongie, S., Carson, C., Greenspan, H. and Malik, J.: Recognition of Images in Large Databases Using a Learning Framework, Technical Report 07-939, UC Berkeley CS Tech Report (1997).
30) Carson, C., Belongie, S., Greenspan, H. and Malik, J.: Region-Based Image Querying, *Proc. IEEE International Workshop on Content-Based Access of Image and Video Libraries* (1997).
31) Carson, C., Belongie, S., Greenspan, H. and Malik, J.: Blobworld: Image Segmentation Using Expectation-Maximization and Its Application to Image Querying, *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol.24, No.8, pp.1026–1038 (2002).
32) Barnard, K. and Forsyth, D.: Learning the Semantics of Words and Pictures, *Proc. IEEE International Conference on Computer Vision*, pp.408–415 (2001).
33) Barnard, K., Duygulu, P., Freitas, N.d., Forsyth, D., Blei, D. and Jordan, M.: Matching Words and Pictures, *Journal of Machine Learning Research*, Vol.3, pp.1107–1135 (2003).
34) Ratan, A.L. and Grimson, W.E.L.: Training templates for scene classification using a few examples, *Proc. IEEE International Workshop on Content-Based Access of Image and Video Libraries*, pp.90–97 (1997).
35) Lipson, P., Grimson, W.E.L. and Sinha, P.: Configuration based scene classification and image indexing, *Proc. IEEE Computer Vision and Pattern Recognition*, pp.1007–1013 (1997).
36) Smith, J.R. and Li, C.S.: Image Classification and Querying Using Composite Region Templates, *Computer Vision and Image Understanding*, Vol.75, No.1/2, pp.165–174 (1999).
37) Maron, O. and Ratan, A.L.: Multiple-instance learning for natural scene classification, *Proc. 15th International Conference on Machine Learning*, pp.341–349 (1998).
38) Ratan, A.L., Maron, O., Grimson, W. and Lozano-Perez, T.: A Framework for Learning Query Concepts in Image Classification, *Proc. IEEE Computer Vision and Pattern Recognition*, pp.423–429 (1999).
39) Dietteric, T.G., Lathro, R.H. and Lozan-Perez, T.: Solving the Multiple Instance Problem with Axis-Parallel Rectangles, *Artificial Intelligence Journal*, Vol.89, pp.31–71 (1997).
40) Duygulu, P., Barnard, K., Freitas, N.d. and Forsyth, D.: Object Recognition as Machine Translation: Learning a Lexicons for a Fixed Image Vocabulary, *Proc. European Conference on Computer Vision*, pp.IV:97–112 (2002).
41) Shi, J. and Malik, J.: Normalized cuts and image segmentation, *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol.22, No.8, pp.888–905 (2000).
42) Brown, P.F., Cocke, J., Pietra, S.D., Pietra, V.D., Jelinek, F., Lafferty, J.D., Mercer, R.L. and Roossin, P.S.: A statistical approach to machine translation, *Computational Linguistic*, Vol.16, No.2, pp.79–85 (1990).
43) Barnard, K., Duygulu, P., Guru, R., Gabbur, P. and Forsyth, D.: The effects of segmentation and feature choice in a translation model of object recognition, *Proc. IEEE Computer Vision and Pattern Recognition*, pp.II: 675–682 (2003).
44) Hofmann, T.: Unsupervised Learning by Probabilistic Latent Semantic Analysis, *Machine Learning*, Vol.43, pp.177–196 (2001).
45) Hofmann, T. and Puzicha, J.: Statistical models for co-occurrence data, Technical Report No.1625, MIT AI Lab. (1998).
46) Mori, Y., Takahashi, H. and Oka, R.: Image-to-word transformation based on dividing and vector quantizing images with words, *Proc. 1st International Workshop on Multimedia Intelligent Storage and Retrieval Management* (1999).
47) 森　靖英, 高橋裕信, 岡　隆一：単語群つき画像の分割クラスタリングによる未知画像からの関連単語推定, 電子情報通信学会論文誌 D-II, Vol.J84-D-II, No.4, pp.649–658 (2001).
48) 森　靖英, 高橋裕信, 保科雅洋, 野崎俊輔, 岡　隆一：WWW 上の文書・画像混在データのクロスメディア検索, 第 15 回情報統合研究会資料 SIG-CII-2001-MAR (2001).
49) Fung, C.Y. and Loe, K.F.: Learning primitive and scene semantics of images for classification and retrieval, *Proc. ACM International Conference Multimedia*, pp.9–12 (1999).
50) Blei, D. and Jordan, M.: Modeling annotated data, *Proc. ACM SIGIR Conference on Research and Development in Information Retrieval*, pp.127–134 (2003).
51) Jeon, J., Lavrenko, V. and Manmatha, R.: Automatic image annotation and retrieval using cross-media relevance models, *Proc. ACM SIGIR Conference on Research and Development in Information Retrieval*, pp.119–126 (2003).
52) Srikanth, M., Varner, J., Bowden, M. and Moldovan, D.: Exploiting ontologies for automatic image annotation, *Proc. ACM SIGIR Conference on Research and Development in*

Information Retrieval, pp.552–558 (2005).
53) Jin, Y., Khan, L., Wang, L. and Awad, M.: Image annotations by combining multiple evidence & wordNet, *Proc. ACM International Conference Multimedia*, pp.706–715 (2005).
54) Kadir, T. and Brady, M.: Scale, Saliency and image description, *International Journal of Computer Vision*, Vol.45, No.2, pp.83–105 (2001).
55) Fergus, R., Perona, P. and Zisserman, A.: Object Class Recognition by Unsupervised Scale-Invariant Learning, *Proc. IEEE Computer Vision and Pattern Recognition*, pp.264–271 (2003).
56) Schmid, C. and Mohr, R.: Local Grayvalue Invariants for Image Retrieval, *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol.19, No.5, pp.530–535 (1997).
57) Harris, C. and Stephens, M.: A Combined Corner and Edge Detector, *Proc. Alvey Conference*, pp.147–152 (1988).
58) Lowe, D.G.: Object recognition from local scale-invariant features, *Proc. IEEE International Conference on Computer Vision*, pp.1150–1157 (1999).
59) Mikolajczyk, K. and Schmid, C.: A performance evaluation of local descriptors, *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol.27, No.10, pp.1615–1630 (2005).
60) Burl, M. and Perona, P.: Recognition of planar object classes, *Proc.IEEE Computer Vision and Pattern Recognition*, pp.223–230 (1996).
61) Burl, M. and Perona, P.: A probabilistic approach to object recognition using local photometry and global geometry, *Proc. European Conference on Computer Vision*, pp.628–641 (1998).
62) Weber, M., Welling, M. and Perona, P.: Towards Automatic Discovery of Object Categories, *Proc. IEEE Computer Vision and Pattern Recognition*, pp.101–108 (2000).
63) Weber, M., Welling, M. and Perona, P.: Unsupervised Learning of Models for Recognition, *Proc. European Conference on Computer Vision*, pp.18–32 (2000).
64) Föstner, W.: A framework for low level feature extraction, *Proc. European Conference on Computer Vision*, pp.383–394 (1994).
65) Fergus, R., Perona, P. and Zisserman, A.: A Sparse Object Category Model for Efficient Learning and Exhaustive Recognition, *Proc. IEEE Computer Vision and Pattern Recognition*, pp.380–387 (2004).
66) Fei-Fei, L., Fergus, R. and Perona, P.: A Bayesian Approach to Unsupervised One-Shot Learning of Object Categories, *Proc. IEEE International Conference on Computer Vision*, pp.1134–1141 (2003).
67) Leibe, B., Leonardis, A. and Schiele, B.: Combined object categorization and segmentation with an implicit shape model, *Proc. ECCV Workshop on Statistical Learning in Computer Vision* (2004).
68) Crandall, D. and Huttenlocher, D.: Weakly supervised learning of part-based spatial models for visual object recognition, *Proc. European Conference on Computer Vision*, pp.I: 16–29 (2006).
69) Ponce, J., Hebert, M., Schmid, C. and Zisserman, A. (Eds.): *Toward Category-Level Object Recognition*, Lecture Note on Computer Science (LNCS), No.4170, Springer-Verlag (2006).
70) Pinz, A.: Object Categorization, *Foundations and Trends in Computer Graphics and Vision*, Vol.1, No.4, pp.255–353 (2006). http://www.emt.tugraz.at/~pinz/onlinepapers/CGV003-journal.pdf
71) Bosch, A., Munoz, X. and Marti, R.: Which is the best way to organize/classify images by contents?, *Image and Vision Computing*, Vol.25, No.6, pp.778–791 (2007).
72) Datta, R., Li, J. and Wang, J.Z.: Content-based image retrieval: Approaches and trends of the new age, *Proc. ACM SIGMM International Workshop on Multimedia Information Retrieval*, pp.253–262 (2005).
73) Csurka, G., Bray, C., Dance, C. and Fan, L.: Visual categorization with bags of keypoints, *Proc. ECCV Workshop on Statistical Learning in Computer Vision*, pp.1–22 (2004).
74) Manning, C.D. and Schütze, H.: *Foundation of Statistical Natural Language Processing*, The MIT Press (1999).
75) Otsu, N. and Kurita, T.: A new scheme for practical flexible and intelligent vision systems, *Proc. IAPR Workshop on Computer Vision*, pp.431–435 (1988).
76) 大津展之，栗田多喜夫，関田　巌：パターン認識—理論と応用，朝倉書店 (1996).
77) Fergus, R., Fei-Fei, L., Perona, P. and Zisserman, A.: Learning Object Categories from Google's Image Search, *Proc. IEEE International Conference on Computer Vision*, pp.1816–1823 (2005).
78) Sivic, J., Russell, B.C., Efros, A.A., Zisserman, A. and Freeman, W.T.: Discovering Objects and their Localization in Images,

Proc. IEEE International Conference on Computer Vision, pp.370–377 (2005).
79) Blei, D., Ng, A. and Jordan, M.: Latent Dirichlet Allocation, *Journal of Machine Learning Research*, Vol.3, pp.993–1022 (2003).
80) Fei-Fei, L. and Perona, P.: A Bayesian Hierarchical Model for Learning Natural Scene Categories, *Proc. IEEE Computer Vision and Pattern Recognition*, pp.524–531 (2005).
81) Deerwester, S.C., Dumais, S.T., Landauer, T.K., Furnas, G.W. and Harshman, R.A.: Indexing by Latent Semantic Analysis, *Journal of the American Society of Information Science*, Vol.41, No.6, pp.391–407 (1990).
82) 上田修功：ベイズ学習，電子情報通信学会誌，Vol.85, No.4, 6, 7, 8 (2002).
83) Lowe, D.G.: Distinctive Image Features from Scale-Invariant Keypoints, *International Journal of Computer Vision*, Vol.60, No.2, pp.91–110 (2004).
84) Sivic, J. and Zisserman, A.: Video Google: A Text Retrieval Approach to Object Matching in Videos, *Proc. IEEE International Conference on Computer Vision*, pp.1470–1477 (2003).
85) Vedalidi, A.: SIFT++. http://vision.ucla.edu/˜vedaldi/code/siftpp/siftpp.html
86) Nowak, E., Jurie, F., Triggs, W. and Vision, M.: Sampling strategies for bag-of-features image classification, *Proc. European Conference on Computer Vision*, pp.IV: 490–503 (2006).
87) Ponce, J., Berg, T., Everingham, M., Forsyth, D., Hebert, M., Lazebnik, S., Marszalek, M., Schmid, C., Williams, C.K.I., Zhang, J. and Zisserman, A.: Dataset Issues in Object Recognition, *in "Toward Category-Level Object Recognition"*, LNCS No.4170, pp.30–50, Springer-Verlag (2006).
88) PASCAL Challenge. http://www.pascal-network.org/challenges/VOC/
89) Jurie, F. and Triggs, B.: Creating Efficient Codebooks for Visual Recognition, *Proc. IEEE International Conference on Computer Vision*, pp.I:604–610 (2005).
90) Comaniciu, D. and Meer, P.: Mean Shift: A Robust Approach toward Feature Space Analysis, *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol.25, No.5, pp.603–619 (2002).
91) Perronnin, F., Dance, C., Csurka, G. and Bressan, M.: Adapted vocabularies for generic visual categorization, *Proc. European Conference on Computer Vision*, pp.IV: 464–475 (2006).
92) Weijer, J.v.d. and Schmid, C.: Coloring local feature extraction, *Proc. European Conference on Computer Vision*, pp.II: 334–348 (2006).
93) ICCV'05 Short Course: Recognizing and Learning Object Categories. http://people.csail.mit.edu/torralba/iccv2005/
94) Zhang, H., Berg, A.C., Maire, M. and Malik, J.: SVM-KNN: Discriminative Nearest Neighbor Classification for Visual Category Recognition, *Proc. IEEE Computer Vision and Pattern Recognition*, pp.2126–2136 (2006).
95) Lazebnik, S., Schmid, C. and Ponce, J.: Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories, *Proc. IEEE Computer Vision and Pattern Recognition*, pp.2169–2178 (2006).
96) Wang, G., Zhang, Y. and Fei-Fei, L.: Using Dependent Regions for Object Categorization in a Generative Framework, *Proc. IEEE Computer Vision and Pattern Recognition*, pp.1597–1604 (2006).
97) Grauman, K. and Darrell, T.: Unsupervised Learning of Categories from Sets of Partially Matching Image Features, *Proc. IEEE Computer Vision and Pattern Recognition*, pp.19–25 (2006).
98) Mutch, J. and Lowe, D.G.: Multiclass Object Recognition with Sparse, Localized Features, *Proc. IEEE Computer Vision and Pattern Recognition*, pp.11–18 (2006).
99) Wolf, L., Bileschi, S. and Meyers, E.: Perception Strategies in Hierarchical Vision Systems, *Proc. IEEE Computer Vision and Pattern Recognition*, pp.2153–2160 (2006).
100) Joachims, T.: SVM^{light}. http://svmlight.joachims.org/
101) Chang, C.C. and Lin, C.J.: *LIBSVM*. http://www.csie.ntu.edu.tw/ cjlin/libsvm/
102) Grauman, K. and Darrell, T.: Pyramid Match Kernels: Discriminative Classification with Sets of Image Features, *Proc. IEEE International Conference on Computer Vision*, pp.1458–1465 (2005). (modified version: MIT-CSAIL-TR-2006-020).
103) Zhang, J., Marszalek, M., Lazebnik, S. and Schmid, C.: Local Features and Kernels for Classification of Texture and Object Categories: A Comprehensive Study, *International Journal of Computer Vision*, Vol.73, No.2, pp.213–238 (2007).
104) Rubner, Y., Tomasi, C. and Guibas, L.J.: The Earth Mover's Distance as a Metric for Image Retrieval, *International Journal of Computer Vision*, Vol.40, No.2, pp.99–121 (2000).
105) Holub, A. and Perona, P.: A Discriminative

Framework for Modelling Object Classes, *Proc. IEEE Computer Vision and Pattern Recognition*, pp.664–671 (2005).
106) Holub, A., Welling, M. and Perona, P.: Combining Generative Models and Fisher Kernels for Object Recognition, *Proc. IEEE International Conference on Computer Vision*, pp.136–143 (2005).
107) Jaakkola, T.S. and Haussler, D.: Exploiting Generative Models in Discriminative Classifiers, *Advances in Neural Information Processing Systems*, pp.487–493 (1999).
108) Berg, A.C., Berg, T.L. and Malik, J.: Shape Matching and Object Recognition Using Low Distortion Correspondences, *Proc. IEEE Computer Vision and Pattern Recognition*, pp.26–33 (2005).
109) Torralba, A., Murphy, K. and Freeman, W.T.: Using the Forest to See the Trees: A Graphical Model Relating Features, Objects and Scenes, *Advances in Neural Information Processing Systems* (2003).
110) Sudderth, E.B., Torralba, A., Freeman, W.T. and Willsky, A.S.: Learning Hierarchical Models of Scenes, Objects, and Parts, *Proc. IEEE International Conference on Computer Vision*, pp.1331–1338 (2005).
111) Kumar, S. and Hebert, M.: A Hierarchical Field Framework for Unified Context-Based Classification, *Proc. IEEE International Conference on Computer Vision*, pp.1284–1291 (2005).
112) Hoiem, D., Efros, A.A. and Hebert, M.: Putting Objects in Perspective, *Proc. IEEE Computer Vision and Pattern Recognition*, pp.2137–2144 (2006).
113) Pearl, J.: *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Morgan Kaufmann (1988).
114) Zhu, Q., Yeh, M.C. and Cheng, K.T.: Multimodal fusion using learned text concepts for image categorization, *Proc. ACM International Conference Multimedia*, pp.211–220 (2006).
115) Boutell, M., Luo, J. and Brown, C.: A generalized temporal context model for classifying image collections, *Multimedia Systems Journal*, Vol.11, No.1, pp.82–92 (2005).
116) Luo, J., Boutell, M. and Brown, C.: Pictures are not taken in a vacuum, *IEEE Signal Processing Magazine*, Vol.23, No.2, pp.101–114 (2006).
117) Boutell, M. and Luo, J.: Bayesian fusion of camera metadata cues in semantic scene classification, *Proc. IEEE Computer Vision and Pattern Recognition* (2004).
118) Fei-Fei, L., Fergus, R. and Perona, P.: Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories, *Proc. IEEE CVPR Workshop of Generative Model Based Vision* (2004).
119) Caltech 101 image dateset. http://www.vision.caltech.edu/Image_Datasets/Caltech101/
120) Bosch, A., Zisserman, A. and Munoz, X.: Image Classification using Random Forests and Ferns, *Proc. IEEE International Conference on Computer Vision* (2007).
121) 堀田一弘：局所領域集合のカーネル主成分分析に基づくカテゴリ識別，画像の認識・理解シンポジウム（MIRU 2007），pp.IS-1–12 (2007).
122) Fei-Fei, L., Fergus, R. and Perona, P.: One-Shot Learning of Object Categories, *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol.28, No.4, pp.594–611 (2006).
123) Caltech 256 image dateset. http://www.vision.caltech.edu/Image_Datasets/Caltech256/
124) Griffin, G., Holub, A. and Perona, P.: Caltech-256 Object Category Dataset, Technical Report 7694, California Institute of Technology (2007).
125) Hanbury, A.: Analysis of keywords used in image understanding tasks, *Proc. International Workshop OntoImage* (2006).
126) TRECVID Home Page. http://www-nlpir.nist.gov/projects/trecvid/
127) 帆足啓一郎，菅野 勝，松本一則：映像情報検索とその評価技術の最前線，情報処理，Vol.46, No.9, pp.1016–1023 (2005).
128) ImageCLEF Home Page. http://ir.shef.ac.uk/imageclef/
129) Petrov, S., Faria, A., Michaillat, P., Berg, A., Klein, D., Malik, J., Faria, A., Stolcke, A. and Stolcke, A.: Detecting Categories in News Video Using Image Features, *Proc. TRECVID Workshop Conference 2006* (2006).
130) Philbin, J., Bosch, A., Chum, O., Geusebroek, J.-M., Sivic, J. and Zisserman, A.: Oxford TRECVID 2006 — Notebook paper, *Proc. TRECVID Workshop Conference 2006* (2006).
131) Volkmer, T., Smith, J.R. and Natsev, A.: A web-based system for collaborative annotation of large image and video collections: An evaluation and user study, *Proc. ACM International Conference Multimedia*, pp.892–901 (2005).
132) Naphade, M., Smith, J., Tesic, J., Chang, S.-F., Hsu, W. and Kennedy, L., Hauptmann,

A. and Curtis, J.: Large-Scale Concept Ontology for Multimedia, *IEEE Trans. Multimedia*, Vol.13, No.3, pp.86–91 (2006).
133) Russell, B.C., Torralba, R., Murphy, K.P. and Freeman, W.T.: LabelMe: A database and web-based tool for image annotation, Technical Report No.2005-025, MIT AI Lab. (2005).
134) LabelMe Project. http://labelme.csail.mit.edu/
135) Ahn, L.v. and Dabbish, L.: Labeling images with a computer game, *Proc. ACM International Conference on Human Factors in Computing Systems* (*CHI*), pp.319–326 (2004).
136) ESP Game. http://www.espgame.org/
137) Google Image Labeler. http://images.google.com/labeler/
138) Peekaboom. http://www.peekaboom.org/
139) Ahn, L.v., Liu, R. and Blum, M.: Peekaboom: A game for locating objects in images, *Proc. ACM International Conference on Human Factors in Computing Systems* (*CHI*), pp.55–64 (2006).
140) Yanai, K.: Generic Image Classification Using Visual Knowledge on the Web, *Proc. ACM International Conference Multimedia*, pp.67–76 (2003).
141) 柳井啓司：一般画像自動分類の実現へ向けた World Wide Web からの画像知識の獲得，人工知能学会論文誌，Vol.19, No.5, pp.429–439 (2004).
142) Wang, J.Z., Li, J. and Wiederhold, G.: SIMPLIcity: Semantics-Sensitive Integrated Matching for Picture Libraries, *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol.23, No.9, pp.947–963 (2001).
143) Fergus, R., Perona, P. and Zisserman, A.: A Visual Category Filter for Google Images, *Proc. European Conference on Computer Vision*, pp.242–255 (2004).
144) Fischler, M. and Bolles, R.: Random sample consensus: A paradigm for model fitting with application to image analysis and automated cartography, *Comm. ACM*, Vol.24, pp.381–395 (1981).
145) Song, X., Lin, C. and Sun, M.: Autonomous visual model building based on image crawling through internet search engines, *Proc. ACM SIGMM International Workshop on Multimedia Information Retrieval*, pp.315–322 (2004).
146) Yahoo 画像検索 API. http://developer.yahoo.co.jp/search/image/V1/imageSearch.html
147) Flickr API. http://www.flickr.com/services/api/
148) Wang, X.-J., Zhang, L., Jing, F. and Ma, W.-Y.: AnnoSearch: Image Auto-Annotation by Search, *Proc. IEEE Computer Vision and Pattern Recognition*, pp.1483–1490 (2006).
149) Angelova, A., Abu-Mostafa, Y. and Perona, P.: Pruning Training Sets for Learning of Object Categories, *Proc. IEEE Computer Vision and Pattern Recognition*, pp.494–501 (2005).
150) Yanai, K. and Barnard, K.: Probabilistic Web Image Gathering, *Proc. ACM SIGMM International Workshop on Multimedia Information Retrieval*, pp.57–64 (2005).
151) 柳井啓司：確率的 Web 画像収集，人工知能学会論文誌，Vol.21, No.1, pp.10–18 (2007).
152) Rosch, E., Mervis, C.B., Gray, W.D., Johnson, D.M. and Boyes-Braem, P.: Basic Objects in Natural Categories, *Cognitive Psychology*, Vol.8, pp.382–439 (1976).
153) Yanai, K. and Barnard, K.: Image Region Entropy: A Measure of "Visualness" of Web Images Associated with One Concept, *Proc. ACM International Conference Multimedia*, pp.420–423 (2005).
154) 柳井啓司，Barnard, K.：一般物体認識のための単語概念の視覚性の分析，情報処理学会論文誌：コンピュータビジョンとイメージメディア，Vol.48, No.SIG10 (CVIM17), pp.88–97 (2007).
155) Palmer, S.E., Rosch, E. and Chase, P.: Canonical Perspective and the perception of objects, *Attention and Performance*, Vol.9, pp.135–151 (1981).
156) 金出武雄：コンピュータビジョンと AI —その関係と無関係，人工知能学会誌，Vol.18, No.3, pp.328–335 (2003).

(平成 19 年 1 月 10 日受付)
(平成 19 年 7 月 31 日採録)

(担当編集委員　和田 俊和)

柳井　啓司（正会員）

1995 年東京大学工学部計数工学科卒業．1997 年東京大学大学院情報工学専攻修士課程修了．1997 年電気通信大学情報工学科助手．2003〜2004 年文部科学省在外研究員として米国アリゾナ大学に滞在．2006 年電気通信大学情報工学科准教授．博士（工学）．一般画像認識，Web からの画像知識獲得等に興味がある．人工知能学会，電子情報通信学会，ACM，IEEE CS 各会員．