社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2006－CVIM－155 (17)
2006／9／8

一般物体認識の現状と今後

柳井 啓司

電気通信大学 情報工学科

「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一つである．人間は数万種類の対象を認識可能であると言われるが，最近まで計算機はわずか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，急速に研究が発展し，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．本稿では一般物体認識の現状と今後の展望について解説を行う．

The Current State and Future Directions on Generic Object Recognition

KEIJI YANAI

Department of Computer Science, The University of Electro-Communications

"Generic object recognition" aims at enabling a computer to recognize objects in images with their category names. It is recognized as one of an ultimate goal of computer vision research. Although human can recognize ten thousands of kinds of objects, it is extremely difficult for a computer to recognize just 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 recognition has been obtained so far. In this paper, we describe the current state and future directions on generic object recognition.

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を **一般物体認識 (generic object recognition)** と呼び，画像認識の研究において最も困難な課題の一つである．

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なわれている．しかしながら，未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言われる[1] 一方で，我々にとっては馴染み深い対象である，例えば「山」「椅子」「ラーメン」についてすら，現状では計算機が画像認識によって，それらが含まれる画像を自動的に特定することは極めて困難である．

一般に実世界画像に対する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 は個々の物体 (the object) を区別する認識であり，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を出力結果とする．一方，classification は物体の種類 (an object) を区別する認識で，人間が決めた分類 (クラス) と画像中の物体を対応付け，物体のクラス名 (多くの場合は一般名称) を出力結果とする．「物体認識」というと identification の認識のことを指すのが一般的であるが，「一般物体認識」は classification の認識を意味する．

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出来るようになった．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，画像の取り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．例えば，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に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，**機械による人間の認識機能の実現**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．

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，本稿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，本稿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．

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盛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言ってよい．今年のコンピュータビジョンに関する最大の会議であるCVPR2006では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CVPR2004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言うことができる．同様の傾向は今年の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いない．そこで，本稿では，一般物体認識について，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．

2. これまでの研究

まず最初に本節では一般物体認識の歴史を述べる．

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年代当初よりの研究が行なわれていた．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，最初に成功を見た研究は，限定された世界『積木の世界』を対象としたものであった．その代表例の線画解釈[3]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，例えば，航空写真などの様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．予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[4]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[5]，The Schema System[6]，SIGMA[7]などの画像中の物体毎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点(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，それを解決するこは出来なかった．

当時の研究のほとんどが3次元画像を航空写真と同じ様に2次元的な画像として取り扱っており，領域分割を行なった後に，関係や構造の情報を利用してそれぞれの領域にラベリングを行い認識を実現していた．この様な方法では，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次元であるのにも拘らず，3次元的な取り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，D. Marrの提案した「視覚認識への計算論的アプローチ」[8]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＋ラベリング規則の様な2次元的な物体認識の手法は下火となった[9]．

その後，3次元の実世界を対象とする認識については，モデルベースト(model-based)による物体認識の研究が盛んに行なわれた[10]．モデルベースト物体認識では，認識の対象とする物体の形状モデルを知識として予め用意しておいて，画像とモデルの照合を行うことにより，画像中にモデルの表す物体の存在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．他にも一般化円筒[11]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なわれた[12]．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を目的としたモデルベースト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る．Identificationには向いているが，classificationに適用することとは困難である．例外的に，プロトタイプモデルによって，モデルベーストアプローチでclassificationを目指した研究[13]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程あり，そのすべての形状が既知であることはあり得ず，また，海や道路などのように明確な形状を定義することできない物体も多く存在する等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[14]，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なうcontext-based recognition[15]，画像エキスパートシステム[16)~18)]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，一般化することは出来なかった．

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．カラーヒストグラム[19]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．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．特徴量が色のみなのでclassification的な物体認識にはシーン認識以外の具体的な物体の認識には向いていないが，特徴量が簡単に求められるので大量データに対するidentificationには極めて有効な手法である[20),21)]．ヒストグラムはテクスチャにも応用された[22)]．

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みなして，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みなす固有顔[23)]が提案された．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[24)]も提案された．これらの方法では，3次元物体を3次元情報を復元せずに2次元の外観(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出来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盛んに研究が行われたとは言えず，「(一般)画像認識の暗黒時代」であった．そうした中で，次に述べる様に，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年代より盛んに研究されている[25),26)]．画像検索は画像認識のコミュニティーではなく，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ーで研究されてきた．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(分類名)が同じであるということである．もし，予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言うことができる．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[27)]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．S. Belongieらによる研究[28),29)]では，Blobworld[30)]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が導入したものが，次節で紹介するword-image translation model である[31),32)]．

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A. L. Ratanらによる研究[33)]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[34)]を拡張したものである．J. R. Smithらによる同様の研究[35)]もある．一方，O. Maronらの研究[36),37)]では，一般物体認識にmultiple instance learning (MIL)[38)]を導入することを提案した．グリッド分割した正サンプルと負サンプルを学習画像として用いて，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特徴を，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IEEE ICME(Inter. Conf. on Multimedia and Expo)などのマルチメディア系(MM)の会議で発表されることが多い．

3. 21世紀の新しい手法：一般物体認識の復活

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，大量のデータの高速処理が可能となったために可能でなったアプローチであると言える．それに伴って，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言うだけではなく，データマイニング (DM)，機械学習 (ML) 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CV系会議，前節で述べたMM系の会議に加えて，NIPS(Neural Information Processing Systems)やICML(Inter. Conf. on Machine Learning)などのML系の会議でも発表される様になってきている．

さて，次に本節では，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ーで行われ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一つの山を越えたと言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word-image translation model[31),32),39)]である．予め画像全体対してに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．Blobworld[30)]もしくはNormalized Cuts[40)]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1)．文単位で対応のとれている2か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[41)]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みなし，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 (image-word 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．

実際には，31)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．森らの研究[43),44)]では，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．この研究は32)で引用されることによって，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．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[45)]もある．他に類似研究として，C. Y. Fungら[46)]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する．そして，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る．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する．

以上述べた研究は，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，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．

Translation modelはICCV2001でオーラルペーパー[31)]として発表されて，さらにECCV2003でbest paper award in cognitive vision[39)]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ーにおいては，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

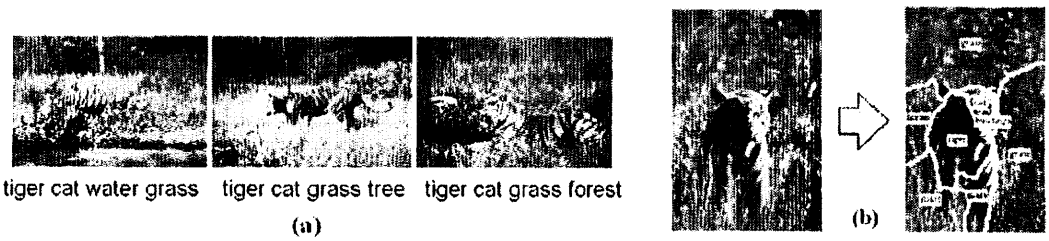

図1 (a) 単語付きの Corel 画像の例. (b) Translation model による画像領域へのアノテーションの例. (図は 42) より引用)

がある. 認識結果を元に領域を統合する試み[47] 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ー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つくことは結果が目でみて分かりやすいからである. 実際, CVPR では translation model の様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[48)~50)].

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行かない場合には, 対処することが難しい. そこで, C. Schmid らは局所的な特徴の組み合わせによって, 画像の照合を行う方法を提案した[53]. 具体的には, 最初に Harris interest point detector[54] によって, 画像中から 100 点程度の特徴点を選び出す. 次に, 各点の画素値や微分値等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にする. 照合は, 未知の画像に対して, 同様に特徴ベクトルの集合を求めて, 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う. 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. この C. 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. D. Lowe も同様の方法によって, 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[55]. 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.

一方, M. C. Burl[56),57)] 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予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 M. Weber らの研究[58),59)] では, constellation model に C. 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[53] 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 interest point detector[60] 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. 局所パターンの表現は特徴点周りの 11×11 の濃淡パターンを用いている. これらの研究の発展させて, 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 R. Fergus らの研究[52] 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[51] 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 2に 52) での「バイク」の認識の例を示す. 図 2(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出来て, クラス内の変動に柔軟に対応できていることが分かる. 61) では 52) の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, L. Fei-Fei ら[62] は, constellation model において, 他のクラスの学習モデルを利用して, 1 枚から 5 枚という僅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ートを占めていたのとは対照的である.

これらの constellation model に関する一連の研究は, すべて P. Perona が率いるカルフォルニア工科大学のグループによって行われている. Constellation model は, 一般物体認識における局所特徴量の有効性を世の中に示したという点でその功績は極めて高いと言える.

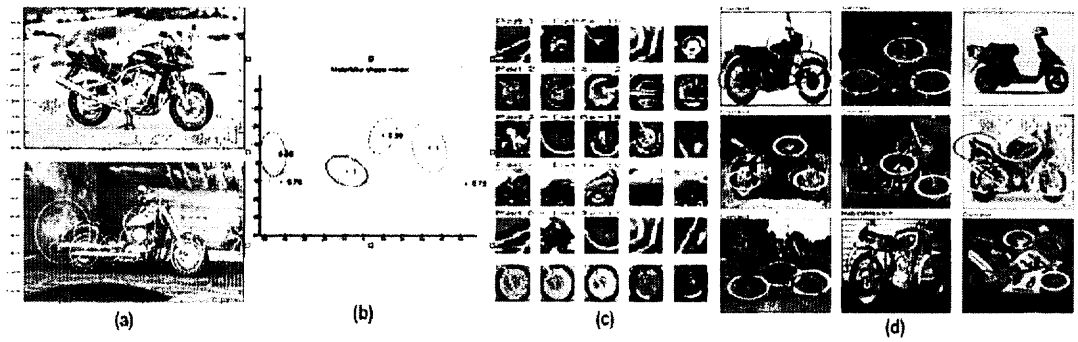

図 2　(a) バイク画像に対する Kadir-Brady detector[51] による検出結果．円の大きさは特徴点のスケールの大きさを示す．(b) 学習された部分の相対位置関係モデル．この例では 6 つの局所特徴から「バイク」モデルを構築している．(c) 局所パターン．この図は 5 枚の認識対象画像から検出された部分画像．(d) 認識結果．(図は 52) より引用)

3.3 最近の研究トピック (1)：bag-of-keypoints

Constellation model では，局所領域の相対的位置の情報も確率モデル化していたが，局所領域の特徴量のみで認識行う方法も提案されており，constellation model に匹敵する認識結果を出している[63)~65)]．こうした位置情報 (shape もしくは geometric information) を使わない方法を bag-of-keypoints approach[63)] と言う．そして，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．

Bag-of-keypoints[63)] は，統計的言語処理における bag-of-words model[66)] のアナロジーで，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) の集合として捉える考え方である．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．

Translation model はまさにテキスト翻訳の手法の応用であるが，実は part-based model にも bag-of-keypoints もしくは visual word の考え方を用いると統計的言語処理の手法を適用することができる．例えば，元々文書分類の手法として提案された probabilistic Latent Semantic Analysis (pLSA)[67)~69)], Latent Dirichlet Allocation (LDA)[70),71)] などが一般物体認識に応用されている．

L. Fei-Fei ら[71)] は，局所パターンを SIFT 特徴量[55),72)] 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) を作成する．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[71)] 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ー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行かないようで，かつて Photobook[27)] などで行われたグリッド分割を用いた結果が最もよい結果になっている点も興味深い．

Bag-of-keypoints では通常，局所特徴間の位置関係は考慮しないが，R. Fergus ら[68)] 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[67)] を物体認識向けに改良した Translation and Scale Invariant pLSA(TSI-pLSA) を提案した．僅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．J. 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[73)]．SIFT 特徴量[55)] 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．

なお，ICCV 2005 でのチュートリアル"Recognizing and Learning Object Categories" のホームページ[74)] 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．

3.4 最近の研究トピック (2)：context

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．例えば，緑の草原の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 context を用いた認識と呼ばれている．Context 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．例えば，context-based recognition[15]，The Schema System[6] 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，context 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 context の利用に再び注目が集まっている．

A. Torralba ら[75] 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．E. B. Sudderth ら[76] や S. Kumar ら[77] 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て，路上シーンや研究室シーンの画像の認識を行った．D. Hoiem ら[78] 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[79] 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．

以上述べたように，context を確率モデルを利用して扱うには，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やベイジアンネットワークを導入するのが一般的で，簡単なシーンの context を表現するのにも大掛かりな仕掛けが必要となっている．さらに対象シーン中に表れる可能性のある物体がそれぞれ認識できる必要もあるので，対象シーンを限定している研究がほとんどで，研究は始まったばかりである．

3.5 最近の研究トピック (3)：生成モデルから判別モデルへ

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 80)~85) のうち，constellation model の研究グループの L. Fei-Fei らの研究[82] 以外はすべて 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．

サポートベクタマシン (SVM) は高い性能を持ったクラス分類手法であるため，part-based アプローチにも導入可能であることが望ましい．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が定義できる事が必要であるが，bag-of-keypoints 表現はベクトルの集合であって，画像によって構成するベクトルの数も異なるので，単一のベクトル同士の時のように簡単に類似度を求めることはできない．そこで，K. Grauman らは Pyramid Match Kernel[64] という多数の画像特徴ベクトルから構成される 2 つの bag 同士の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．S. Lazebnik ら[81] は，Pyramid Match Kernel[64] に局所特徴の位置も考慮するように改良を加えた Spatial Matching を提案した．

一方，constellation model に SVM を取り入れる研究[86),87] も行われている．87) では，generative approach を discriminative approach に変換するための Fisher kernel[88] を constellation model に適用し，SVM を用いた分類を行った．実験では，従来の generative な方法[52] に対して性能向上が見られた．

H. Zhang ら[80] は，最近傍分類法 (Nearest Neighbor) と SVM を組合せた SVM-KNN 法を提案した．SVM-KNN は簡単に言うと，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よ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[89] 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[89] 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，今回のテーマセッションのテーマである「事例ベース」的な判別手法に変化しつつある．

4. 評価の方法 と グランドツルースの作成

本節では，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．

4.1 評価データセット

以上述べたように研究が盛んになると，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[32] 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取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訳ではないので，対象によって認識の難易度の差が大き過ぎるという問題点が

表1 Caltech 101 の平均分類精度

順位	グループ	発表学会	文献 no.	結果 (%)
1	UCB	CVPR'06	80)	**66.23**
2	INRIA	CVPR'06	81)	**64.6**
3	UIUC	CVPR'06	82)	**63**
4	MIT	ICCV'05	64)	**58**
5	UBC	CVPR'06	84)	**56**
6	MIT	CVPR'06	85)	**51.2**
7	U Amsterdam	CVPR'06 WS	93)	**42.3**
参考	Caltech	PAMI'06	92)	**17.7**

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[90),91)] 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．Web 上[91)] 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，その名の通り 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 9144 枚の画像から構成される．クラス毎に枚数が異なり，31 枚から 800 枚までとばらつきがある．Airplane から zebra まで人間の正面顔画像 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．52)，62) 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 2(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．

現在，Caltech101 を使った画像分類において，最もよい結果を出しているのが UC Berkeley のグループが認識率 66.23%[80)] である．これは各クラス毎にランダムに 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に今年の CVPR2006 までに発表された上位 7 位までの結果を示す．データセットが公開されたのが 2004 年のため，2005 年か 2006 年の結果のみであるが，今後も年々上位の結果は向上するものと思われる．なお，3 位の 82) 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%[92)] であり，ここの 2, 3 年のレベルアップは目覚ましい．

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．例えば，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極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 をいれておくと全体の認識率の数値を上げることが出来るからである．そうした問題を解決するには，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 Caltech-101[90)] 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多いに貢献していると言える．

4.2 ベンチマークワークショップ

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，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，コミュニティー全体の技術を向上させていくことが目的である．PASCAL Challenge[94)], TRECVID[95),96)], ImageCLEF[97)] 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挙げられる．

PASCAL Challenge[94)] はヨーロッパ画像認識コミュニティーの PASCAL によって主催されているコンテストで，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易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 2800 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．

TRECVID[95),96)] は，アメリカの国立の技術標準化機関 NIST(National Institute of 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，例えば，対象の explosion, car 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．他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 2000 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 (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，時間の掛かる認識手法は困難である．2006 年度は 14 万ショット以上に対して 39 種類すべてについて認識を行わないといけないため，参加のための敷居が高いという問題がある．

認識結果は 2005 年の 10 課題 (people_walking, explosion, map, US_flag, building, waterscape, mountain, prisoner, sport, car) の場合，平均で 4 割程度であった．表 1中の 7 位の U Amsterdam のグループは TRECVID にも参加して，提案手法を TRECVID にも適用しており，一般物体認識の研究が映像検索にも適用できることを示すものとなっている．TRECVID は映像が対象のため，参加者の多くは一般物体認識の研究者ではなく，映像処理の研究者である．

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 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[97] 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 1000 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いるが，情報検索の研究者の参加が中心になっている．

4.3 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出来たが，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 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．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掛けて集めることが必要である．

Caltech-101 は 9000 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[98]．また，TRECVID のニュース映像データについて，人手によって 1000 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)[99] というプロジェクトもある．1000 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ーを考慮して利用価値の高い 1000 種類のコンセプトの定義を目指している．

他に，大まかに領域分割された画像の全部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[100),101)] 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掛かり，画像中からの物体の位置の検出まで含めた認識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[102]．EPS game[103] は，CMU の学生の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．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 1000 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．1 枚の画像に対して複数のプレーヤー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[102]．

4.4 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[104),105)] 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) の研究が盛んに行われているが，104), 105) 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 から自動的に獲得する．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 Distance(EMD)[106] および Integrated Region Matching(IRM)[107] 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[52] の提案者 R. Fergus らは，Google Image Search の結果の画像からを用いて認識モデルの学習[68),108] を行った．この研究では，Google Image Search の出力から RANSAC[109] 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．他にも，Web 上の画像を用いた物体認識の研究[110] 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．

一方，AnnoSreach[111] 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．104), 105) では，予め分類クラスを指定する必要があったが，111) ではその必要はなく，どの様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，常に誤った知識 (ノイズ) が含まれている．例えば，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．R. Fergus ら[108] はモデル学習時に RANSAC[109] を用いた．一方，A. Angelova ら[112] 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．K. Yanai[113] は EM アルゴリズムを応用した繰り返し手法によって，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．

68) 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ヶ国語に翻訳し，7ヶ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常にノイズを含むために，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．

5. 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．

例えば，次の様な問題が考えられる．

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

本節ではそれぞれについて述べる．

5.1 多種類化 と 認識クラスの決め方

今後は多種類へ対応はますます進んで，1, 2 年以内には 1000 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)[99] プロジェクトでは，1000 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．我々も現在，1000 種類のカテゴリーの画像を Web から各 1000 枚ずつ収集を行っているが，1000 種類になるとカテゴリー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た．それは，一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言われる[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」でもあるかも知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」でもあるとも言え，また，made-of 関係を考えると「乗用車」は「鉄板」や「ゴム」「ガラス」などであるとも言える．

このため，言語の階層構造を常に考慮して，認識すべきクラスに対応する名詞を選ぶ必要がある．E. Rosch ら[114]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，人がぱっと見た時にすぐに思いついたり，幼児が最初に覚えるような，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易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，単語が表わす「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[42),115)]．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．

5.2 クラス内変化への対応

Caltech-101 は種類は 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．パラメトリック固有空間法[24)] の classification 版が実現できれば，有効性が高いであろう．

こうした人間にとって典型的な見え方を canonical perspective[116)] といい，116) 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” 認識が出来たと認めることが出来るであろう．

ただし，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しらべるのは困難である．Web などから同一クラスの大量の画像を用意して，このクラスはこのビューの画像が典型的というのを，クラスタリングなどによって自動的に探し出すことは可能であろうが，そうした研究は行われていない．

また，クラスによってはクラス内での変動が大きい場合もある．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．例えば，「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，ソファーの様な椅子，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．

他に，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である程度可能であるが，かなりの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 context 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 context 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，context には，物体同士の物理的な関係 (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．

6. おわりに

本稿では，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おこなっているが，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" が実現できたと言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，context を含めた様々な種類の知識を大量のデータ，特に Web 上にある情報から獲得することが実現のための鍵となると考えている．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留まらず，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[117)]．

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) Clowes, M.B.: On Seeing things, *Artificial Intelligence*, Vol.2, No.1, pp.79–116 (1971).
4) Tenenbaum, J.M. and Barrow, H.G.: Experiments in Interpretation Guided Segmentation, *Artificial Intelligence*, Vol.8, pp.241–274 (1977).
5) Ohta, Y.: *Knowledge-Based Interpretation of Outdoor Natural Color Scenes*, Pitman Advanced Publishing Program, Boston (1985).
6) 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).
7) Matsuyama, T. and Hwang, V.S.: *SIGMA: A knowledge-based aerial image understanding system*, Plenum Press, New York (1990).
8) Marr, D.: *Vision*, Freeman (1982). (乾, 安藤 (訳): ビジョン, 産業図書 (1985)).
9) 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).
10) 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).
11) Binford, T.: Visual Perception by Computer, *Proc. of IEEE Conf. on Systems and Control* (1975).
12) Brooks, R.A.: Model-Based Three-Dimensional Interpretations of Two-Dimensional Image, *IEEE Trans. on PAMI*, Vol.5, No.2, pp.140–150 (1983).
13) Basri, R.: Recognition by Prototypes, *International Journal of Computer Vision*, Vol.10, No.2, pp.147–167 (1996).
14) Stark, L. and Bowyer, K.: Achieving Generalized Object Recognition through Reasoning about Association of Function to Structure, *IEEE Trans. on PAMI*, Vol.13, No.10, pp.1097–1104 (1991).
15) Strat, T.M. and Fischler, M.A.: Context-Based Vision: Recognizing Objects Using Information from Both 2-D and 3-D Imagery, *IEEE Trans. on PAMI*, Vol.13, No.10, pp.1050–1065 (1991).
16) 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).
17) 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).
18) 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).
19) Swain, M.J. and Ballard, D.H.: Color Indexing, *International Journal of Computer Vision*, Vol.7, No.1, pp.11–32 (1991).
20) 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).
21) Kashino, K., Kurozumi, T. and Murase, H.: A Quick Search Method for Audio and Video Signals Based on Histogram Pruning, *IEEE Transaction on Multimedia*, Vol.5, No.3, pp.348–357 (2003).
22) Schiele, B. and Crowley, J.L.: Recognition using Multidimensional Receptive Field Histograms, *ECCV*, pp.610–619 (1996).
23) Turk, M. and Pentland, A.: Eigenfaces for Recognition, *Cognitive Neuroscience*, Vol.3, No.1, pp.71–96 (1991).
24) 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).
25) Gudivada, V.N. and Raghavan, V.V.: Content-Based Image Retrieval-Systems, *IEEE Computer*, Vol.28, No.9, pp.18–22 (1995).
26) 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, Vol.40, No.SIG3(TOD 1), pp.171–184 (1999).
27) Minka, T.P. and Picard, R.W.: Vision Texture for Annotation, *ACM/Springer Journal of Multimedia Systems*, Vol.3, pp.3–14 (1995).
28) Belongie, S., Carson, C., Greenspan, H. and Malik, J.: Recognition of Images in Large Databases Using a Learning Framework, Technical Report07-939, UC Berkeley CS Tech Report (1997).
29) Carson, C., Belongie, S., Greenspan, H. and Malik, J.: Region-Based Image Querying, *Proc. of IEEE International Workshop on Content-Based Access of Image and Video Libraries* (1997).
30) Carson, C., Belongie, S., Greenspan, H. and Malik, J.: Blobworld: Image Segmentation Using Expectation-Maximization and Its Application to Image Querying, *IEEE Trans. on PAMI*, Vol.24, No.8, pp.1026–1038 (2002).
31) Barnard, K. and Forsyth, D.A.: Learning the Semantics of Words and Pictures, *ICCV*, pp.408–415 (2001).
32) Barnard, K., Duygulu, P., deFreitas, N., Forsyth, D., Blei, D. and Jordan, M.: Matching Words and Pictures, *Journal of Machine Learning Research*, Vol.3, pp.1107–1135 (2003).
33) Ratan, A.L. and Grimson, W.E.L.: Training templates for scene classification using a few examples, *Proc. of IEEE International Workshop on Content-Based Access of Image and Video Libraries*, pp.90–97 (1997).
34) Lipson, P., Grimson, W.E.L. and Sinha, P.: Configuration based scene classification and image indexing, *CVPR*, pp.1007–1013 (1997).
35) 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).
36) Maron, O. and Ratan, A.L.: Multiple-instance learning for natural scene classification, *Proc. of 15th International Conference on Machine Learning*, pp.341–349 (1998).
37) Ratan, A.L., Maron, O., Grimson, W. and Lozano-Perez, T.: A Framework for Learning Query Concepts in Image Classification, *CVPR*, pp.423–429 (1999).
38) 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).

39) Duygulu, P., Barnard, K., de Freitas, J. F. G. and Forsyth, D.A.: Object Recognition as Machine Translation: Learning a Lexicons for a Fixed Image Vocabulary, *ECCV*, pp.IV:97–112 (2002).

40) Shi, J. and Malik, J.: Normalized cuts and image segmentation, *IEEE Trans. on PAMI*, Vol.22, No.8, pp.888–905 (2000).

41) Brown, P., Cocke, J., DellaPietra, S., DellaPietra, V., Jelinek, F., Lafferty, J., Mercer, R. and Roossin, P.: A statistical approach to machine translation, *Computational Linguistic*, Vol.16, No.2, pp.79–85 (2000).

42) 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, 情報処理学会コンピュータビジョン・イメージメディア研究会報告 CVIM2005-152-1, pp. 1–8 (2006).

43) Mori, Y., Takahashi, H. and Oka, R.: Image-to-word transformation based on dividing and vector quantizing images with words, *Proc. of First International Workshop on Multimedia Intelligent Storage and Retrieval Management* (1999).

44) 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).

45) 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).

46) Fung, C.Y. and Loe, K.F.: Learning primitive and scene semantics of images for classification and retrieval, *ACM Multimedia*, pp.9–12 (1999).

47) 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, *CVPR*, pp.II:675–682 (2003).

48) Jeon, J., Lavrenko, V. and Manmatha, R.: Automatic image annotation and retrieval using cross-media relevance models, *ACM SIGIR*, pp.119–126 (2003).

49) Srikanth, M., Varner, J., Bowden, M. and Moldovan, D.: Exploiting ontologies for automatic image annotation, *ACM SIGIR*, pp. 552–558 (2005).

50) Jin, Y., Khan, L., Wang, L. and Awad, M.: Image annotations by combining multiple evidence & wordNet, *ACM Multimedia*, pp.706–715 (2005).

51) Kadir, T. and Brady, M.: Scale, Saliency and image description, *International Journal of Computer Vision*, Vol.45, No.2, pp.83–105 (2001).

52) Fergus, R., Perona, P. and Zisserman, A.: Object Class Recognition by Unsupervised Scale-Invariant Learning, *CVPR*, pp.264–271 (2003).

53) Schmid, C. and Mohr, R.: Local Grayvalue Invariants for Image Retrieval, *IEEE Trans. on PAMI*, Vol.19, No.5, pp.530–535 (1997).

54) Harris, C. and Stephens, M.: A Combined Corner and Edge Detector, *Proc. of Alvey Conference*, pp. 147–152 (1988).

55) Lowe, D.G.: Object recognition from local scale-invariant features, *ICCV*, pp.1150–1157 (1999).

56) Burl, M. and Perona, P.: Recognition of planar object classes, *CVPR*, pp.223–230 (1996).

57) Burl, M. and Perona, P.: A probabilistic approach to object recognition using local photometry and global geometry, *ECCV*, pp.628–641 (1998).

58) Weber, M., Welling, M. and Perona, P.: Towards Automatic Discovery of Object Categories, *CVPR*, pp.101–108 (2000).

59) Weber, M., Welling, M. and Perona, P.: Unsupervised Learning of Models for Recognition, *ECCV*, pp.18–32 (2000).

60) Föstner, W.: a framework for low level feature extraction, *ECCV*, pp.383–394 (1994).

61) Fergus, R., Perona, P. and Zisserman, A.: A Sparse Object Category Model for Efficient Learning and Exhaustive Recognition, *CVPR*, pp. 380–387 (2004).

62) Fei-Fei, L., Fergus, R. and Perona, P.: A Bayesian Approach to Unsupervised One-Shot Learning of Object Categories, *ICCV*, pp.1134–1141 (2003).

63) Csurka, G., Bray, C., Dance, C. and Fan, L.: Visual categorization with bags of keypoints, *Proc. of ECCV Workshop on Statistical Learning in Computer Vision*, pp.1–22 (2004).

64) Grauman, K. and Darrell, T.: Pyramid Match Kernels: Discriminative Classification with Sets of Image Features, *ICCV*, pp. 1458–1465 (2005). (modified version: MIT-CSAIL-TR-2006-020).

65) Opelt, A., Pinz, A. and Zisserman, A.: Incremental learning of object detectors using a visual shape alphabet, *CVPR*, pp.3–10 (2006).

66) Manning, C.D. and SchFütze, H.: *Foundation of Statistical Natural Language Processing*, The MIT Press (1999).

67) Hofmann, T.: Unsupervised Learning by Probabilistic Latent Semantic Analysis, *Machine Learning*, Vol.43, pp.177–196 (2001).

68) Fergus, R., Fei-Fei, L., Perona, P. and Zisserman, A.: Learning Object Categories from Google's Image Search, *ICCV*, pp.1816–1823 (2005).

69) Sivic, J., Russell, B.C., Efros, A.A., Zisserman, A. and Freeman, W.T.: Discovering Objects and their Localization in Images, *ICCV*, pp.370–377 (2005).

70) Blei, D., Ng, A. and Jordan, M.: Latent Dirichlet Allocation, *Journal of Machine Learning Research*, Vol.3, pp.993–1022 (2003).

71) Fei-Fei, L. and Perona, P.: A Bayesian Hierarchical Model for Learning Natural Scene Categories, *CVPR*, pp.524–531 (2005).

72) Lowe, D. G.: Distinctive Image Features from Scale-Invariant Keypoints, *International Journal of Computer Vision*, Vol. 60, No. 2, pp. 91–110 (2004).

73) Sivic, J. and Zisserman, A.: Video Google: A Text Retrieval Approach to Object Matching in Videos, *ICCV*, pp.1470–1477 (2003).

74) ICCV'05 Short Course: Recognizing and Learning Object Categories: *http://people.csail.mit.edu/torralba/iccv2005/*.

75) Torralba, A., Murphy, K. and Freeman, W.: Using the Forest to See the Trees: A Graphical Model Relating Features, Objects and Scenes, *Advances in Neural Information Processing Systems* (2003).

76) Sudderth, E. B., Torralba, A., Freeman, W. T. and Willsky, A.S.: Learning Hierarchical Models of Scenes, Objects, and Parts, *ICCV*, pp.1331–1338 (2005).

77) Kumar, S. and Hebert, M.: A Hierarchical Field Framework for Unified Context-Based Classification, *ICCV*, pp.1284–1291 (2005).

78) Hoiem, D., Efros, A.A. and Hebert, M.: Putting Objects in Perspective, *CVPR*, pp. 2137–2144 (2006).
79) Pearl, J.: *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Morgan Kaufmann (1988).
80) Zhang, H., Berg, A.C., Maire, M. and Malik, J.: SVM-KNN: Discriminative Nearest Neighbor Classification for Visual Category Recognition, *CVPR*, pp.2126–2136 (2006).
81) Lazebnik, S., Schmid, C. and Ponce, J.: Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories, *CVPR*, pp.2169–2178 (2006).
82) Wang, G., Zhang, Y. and Fei-Fei, L.: Using Dependent Regions for Object Categorization in a Generative Framework, *CVPR*, pp.1597–1604 (2006).
83) Grauman, K. and Darrell, T.: Unsupervised Learning of Categories from Sets of Partially Matching Image Features, *CVPR*, pp.19–25 (2006).
84) Mutch, J. and Lowe, D. G.: Multiclass Object Recognition with Sparse, Localized Features, *CVPR*, pp.11–18 (2006).
85) Wolf, L., Bileschi, S. and Meyers, E.: Perception Strategies in Hierarchical Vision Systems, *CVPR*, pp.2153–2160 (2006).
86) Holub, A. and Perona, P.: A Discriminative Framework for Modelling Object Classes, *CVPR*, pp.664–671 (2005).
87) Holub, A., Welling, M. and Perona, P.: Combining Generative Models and Fisher Kernels for Object Recognition, *ICCV*, pp.136–143 (2005).
88) Jaakkola, T.S. and Haussler, D.: Exploiting Generative Models in Discriminative Classifiers, *Advances in Neural Information Processing Systems*, pp.487–493 (1999).
89) Berg, A. C., Berg, T. L. and Malik, J.: Shape Matching and Object Recognition Using Low Distortion Correspondences, *CVPR*, pp.26–33 (2005).
90) 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. of IEEE CVPR Workshop of Generative Model Based Vision* (2004).
91) Caltech 101 image dateset: *http://www.vision.caltech.edu/Image_Datasets/Caltech101/Caltech101.html*.
92) Fei-Fei, L., Fergus, R. and Perona, P.: One-Shot Learning of Object Categories, *IEEE Trans. on PAMI*, Vol.28, No.4, pp.594–611 (2006).
93) Germert, J. C.v., Geusebroek, J.M., Veenman, C.J., Snoek, C. G.M. and Smeulders, A. W.M.: Robust Scene Categorization by Learning Image Statistics in Context, *Proc. of IEEE CVPR Workshop on Semantic Learning Applications in Multimedia* (2006).
94) PASCAL Challenge: *http://www.pascal-network.org/challenges/VOC/*.
95) TRECVID Home Page: *http://www-nlpir.nist.gov/projects/trecvid/*.
96) 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).
97) ImageCLEF Home Page: *http://ir.shef.ac.uk/imageclef/*.
98) 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, *ACM Multimedia*, pp.892–901 (2005).
99) Naphade, M., Smith, J., Tesic, J., Chang, S.-F., Hsu, W.andKennedy, L., Hauptmann, A. and Curtis, J.: Large-Scale Concept Ontology for Multimedia, *IEEE Transaction on Multimedia*, Vol.13, No.3, pp.86–91 (2006).
100) Russell, B. C., Torralba, R., Murphy, K. P. and Freeman, W. T.: LabelMe: a database and web-based tool for image annotation, Technical Report Memo No.2005-025, MIT AI Lab. (2005).
101) LabelMe Project: *http://labelme.csail.mit.edu/*.
102) Ahn, L.v. and Dabbish, L.: Labeling images with a computer game, *Proc. of ACM International Conference on Human Factors in Computing Systems (CHI)*, pp.319–326 (2004).
103) EPS Game: *http://www.epsgame.org/*.
104) Yanai, K.: Generic Image Classification Using Visual Knowledge on the Web, *ACM Multimedia*, pp. 67–76 (2003).
105) 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, 人工知能学会誌, Vol.19, No.5, pp.429–439 (2004).
106) 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).
107) Wang, J.Z., Li, J. and Wiederhold, G.: SIMPLIcity: Semantics-Sensitive Integrated Matching for Picture Libraries, *IEEE Trans. on PAMI*, Vol.23, No.9, pp.947–963 (2001).
108) Fergus, R., Perona, P. and Zisserman, A.: A Visual Category Filter for Google Images, *ECCV*, pp. 242–255 (2004).
109) Fischler, M. and Bolles, R.: Random sample consensus: A paradigm for model fitting with application to image analysis and automated cartography, *Communications of the ACM*, Vol.24, pp.381–395 (1981).
110) Song, X., Lin, C. and Sun, M.: Autonomous visual model building based on image crawling through internet search engines, *ACM SIGMM WS Multimedia Information Retrieval*, pp.315–322 (2004).
111) Wang, X.-J., Zhang, L., Jing, F. and Ma, W.-Y.: AnnoSearch: Image Auto-Annotation by Search, *CVPR*, pp.1483–1490 (2006).
112) Angelova, A., Abu-Mostafa, Y. and Perona, P.: Pruning Training Sets for Learning of Object Categories, *CVPR*, pp.494–501 (2005).
113) Yanai, K. and Barnard, K.: Probabilistic Web Image Gathering, *ACM SIGMM WS Multimedia Information Retrieval*, pp.57–64 (2005).
114) 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).
115) Yanai, K. and Barnard, K.: Image Region Entropy: A Measure of "Visualness" of Web Images Associated with One Concept, *ACM Multimedia*, pp.420–423 (2005).
116) Palmer, S.E., Rosch, E. and Chase, P.: Cannonical Perspective and the perception of objects, *Attention and Performance*, Vol.9, pp.135–151 (1981).
117) 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).