社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

[招待講演] 一般物体認識における機械学習の利用

柳井　啓司†

† 電気通信大学 大学院情報理工学研究科 総合情報学専攻
〒 182–8585 東京都調布市調布ヶ丘 1–5–1
E-mail: yanai@cs.uec.ac.jp

あらまし　近年，画像認識研究において「一般物体認識」が注目を集めている．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，画像認識研究の究極の研究課題の一つである．2000 年以前はその困難性のため，ほとんど研究が行われていなかったが，(1) 新しい画像表現の提案，(2) 機械学習手法の進歩，(3)Web の普及によるデータセット構築の容易化，(4) 計算機の高速化・大容量化，によって，近年研究が大きく発展し，一般物体認識の実用化が現実のものとなってきている．本稿では，まず，一般物体認識が可能となった最も大きな要因である (1) の新しい画像表現について解説し，さらに (2) の機械学習との関係について焦点を当て，近年の機械学習手法の進歩が一般物体認識の発展に大きく貢献していることを解説する．
キーワード　一般物体認識, 画像表現，機械学習，大量データ

[Invited Talk] Machine Learning in Object Recognition

Keiji YANAI†

† Department of Informatics, Graduate School of Informatics and Engineering,
The University of Electro-Communications　v
Chofugaoka 1–5–1, Chofu-shi, Tokyo, 182–8585 Japan
E-mail: yanai@cs.uec.ac.jp

Abstract　"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 treated in generic object recognition have broad variability regarding their appearance, which makes the problem very tough. For these several years, due to proposal of novel representation of visual models, progress of machine learning methods, facilitation of building of a large-scale data sets, and speeding-up of computers, research on generic object recognition has progressed greatly. In this report, we explain how machine learning techniques play an important role in generic object recognition.
Key words　generic object recognition, image representation, machine learning, large amounts of visual data

1. はじめに

デジタルカメラの普及と，記録メディア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写真を撮影し，それを蓄積することが容易に可能となった．しかしながら，現状では，写真が表すシーンや対象物の意味内容を計算機が理解することができないため，撮影日時などの撮影時に自動的に記録される付加情報（メタデータ）に基づいてデジタル写真の分類や検索が行われることが一般的である．

そうした状況に対して，研究レベルにおいては，写真画像中の「山」「椅子」「ライオン」などの一般的なシーンの意味カテゴリーや，画像に含まれる物体の意味カテゴリーを認識する**一般物体認識 (generic object recognition)** の研究が近年活発に行われるようになって来ている [1].

一般の人が撮影したデジタル写真は，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もしくはシーンを一般的な名称で認識することを「一般物体認識」と呼び，画像認識の研究において最も困難な課題の一つとされている．なぜなら，制約のない画像における「一般的な名称」が表す同一カテゴリーの範囲が広く，同一カテゴリーに属する対象の見た目の変化が極めて大きいために，(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

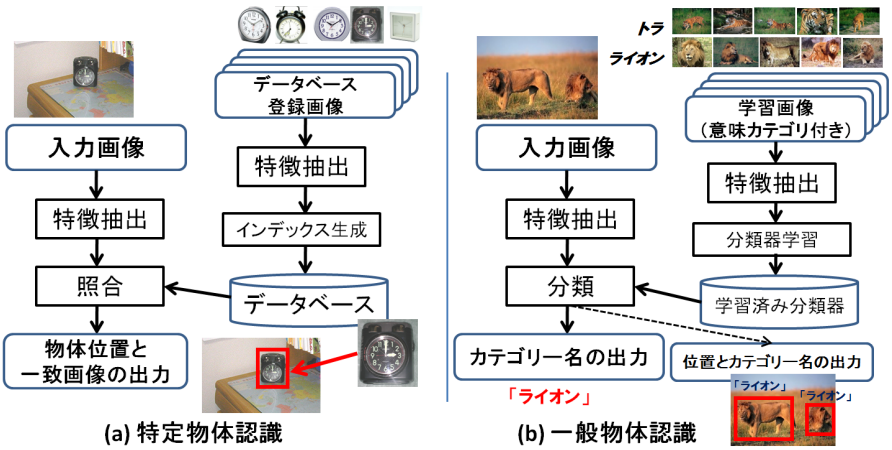

Figure 1　特定物体認識と一般物体認識の違い．同一物体をデータベース中から検索するのが特定物体認識，意味カテゴリーを当てるのが一般物体認識．一般物体認識では，位置検出も行うと一段難しいタスクになる．

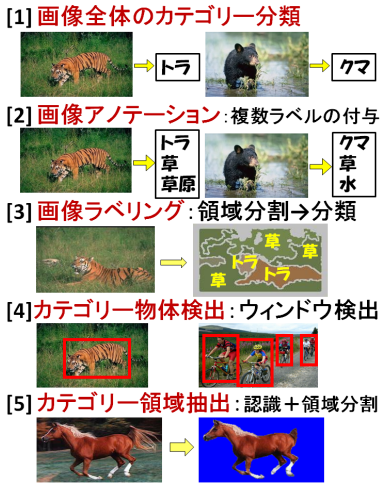

Figure 2　一般物体認識 (カテゴリー認識) の主要な 5 種類のタスク．

の構築，が困難なためである．そのため，2000 年以前はほとんど研究が行われていなかったが，近年の，(1) 新しい画像表現の提案，(2) 機械学習手法の進歩，(3)Web の普及によるデータセット構築の容易化，(4) 計算機の高速化・大容量化，によって研究が大きく発展し，一般物体認識の実用化が現実のものとなってきている．

それに加えて基本的なソフトウェアやデータセットがすべてネット上で公開されているために，研究に参入する敷居が極めて低くなっているという状況がある．それにより多くの研究者が参入し，現在，画像認識の研究コミュニティーでは一般物体認識ブームにあると言ってよい状況になっており，それが研究をますます活性化していると言える．以前は，画像認識の研究者以外が画像認識の実験やシステムを実装することを行うことは容易ではなく，機械学習の研究者が学習アルゴリズムの実験のために画像データを用いることは敷居が高い状況であったが，現在は Web 上から入手可能な特徴抽出ツールや標準データセットによって，機械学習の研究者も気軽に一般物体認識を学習アルゴリズムのテストベッドとして利用できる状況になっている．

本稿では，まず一般物体認識の基本的な画像表現である bag-of-features について解説した後，機械学習との関係について焦点を当て，近年の機械学習手法の進歩が一般物体認識の発展に大きく貢献していることを解説する．

2.　一般物体認識とは？

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．

一方，同じ画像のカテゴリー分類である，文字認識や顔検出・認識は，その実用的な価値から画像認識研究の当初より現在まで研究が続けられてきており，郵便番号の自動読み取りやデジタルカメラの顔認識などで，すでに実用的に使われる段階になっている．こうした，文字や顔の認識は「パターン認識」として，特徴抽出法と特徴分類手法が長い間研究されて，機械学習分野とは関連が深い．

現在，物体認識の研究は，大きく分けて，画像中の物体のカテゴリーを認識する**一般物体認識**と，画像中の個別の物体の認識する**特定物体認識**の 2 通りの研究が行われている．

一般物体認識は，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もしくはシーンを「山」「ライオン」「ラーメン」などの一般的な名称で認識することで，画像認識の研究において最も困難な課題の一つとされている．なぜなら，制約のない画像における「一般的な名称」が表す同一カテゴリーの範囲が広く，同一カテゴリーに属する対象の見た目の変化が極めて大きいために，(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．特に (3) は一般物体認識で固有の問題で，厳密に定義することが不可能な「山」「ライオン」などの意味カテゴリーをいかに定義するかという問題に関係していて，人工知能の分野とも関係の深い問題である．

一方，**特定物体認識**は，「東京タワー」などの特定のランドマークや「iPhone」などの特定の工業製品のようなまったく同じ形状の物体に対する認識技術で，一般物体認識の困難点「(1) 対象の特徴抽出」はほぼ同様であるが，「(2) 認識モデルの構築」は代わりに大量の画像データベースに対して高速な検索を行うことが研究課題となっている．「(3) データセットの構築」の問題は，特定物体認識ではまったく同一のものを探すのが目的であるので，カテゴリーの定義に関する問題は存在しない．

図 1 に 2 つの認識ついての処理の流れについて記す．特定物体認識では，例えば，多くの時計の写真をデータベースに登録しておいて，同一の時計が入力画像に存在するかを調べる．認識対象の画像中の局所パターンとほぼ一致する局所パターンをもつ画像をデータベース中から検索することによって認識を行うため，物体の位置検出も同時に行うことができる．一方，一

般物体認識の例では，入力画像が「ライオン」か「トラ」かを判定するが，その際に入力画像とまったく同じライオンの写真が学習画像になくても，学習による汎化によって，それがライオンであると認識する必要があり，特定物体認識より困難な問題であると言える．特定物体認識と異なり，局所パターンの直接の対応でなく，その分布を用いて認識を行うため，位置まで特定する場合はさらに一段難しい問題となる．

図 2 に参考までに現在研究されている一般物体認識の主なタスクを 5 種類示す．**画像全体のカテゴリー分類**が最も標準的なタスクで，複数のカテゴリーラベルを画像に付与する**画像アノテーション**，領域分割された画像の各領域に対してカテゴリーラベルを付与する**画像ラベリング**，長方形の矩形で画像中の物体の存在位置を検出する**カテゴリー物体検出**, 物体の領域を正確に切り出す**カテゴリー領域抽出**などのタスクが研究課題として扱われている．

以上のように一般物体認識と特定物体認識は，その目的や処理や認識モデルは異なっているが，最も基礎となる認識対象の特徴表現手法はほぼ同一である．どちらも，画像から多数の対象に特徴的な局所的なパターンを局所特徴量として抽出して認識に利用する．次節では，その基本的な手法について解説する．

一般物体認識は，例えば標準データセットの 1 つである 256 種類のカテゴリーから成る Caltech-256 画像データセットの分類精度が 50%程度に留まっていることからも分かるように，まだ実用化するには早い段階であると言える．Caltech-256 は 1 枚の画像に 1 つの物体のみが含まれているという制約があるが，1 枚の画像に複数の物体が含まれていている場合は，物体の切り出しも行う必要があるため，さらに精度は低下し，20 種類の標準的なデータセットに対して最新の結果でも 30%程度の精度に留まっている．これは，すでにデジタルカメラに搭載され実用化されている顔検出のほぼ 100%に近い精度に比べると極めて低い精度である．また，一般物体認識は，画像中の物体や画像の表すシーンを一般名称に対応するカテゴリーに分類する問題であるために，カテゴリーの定義自体がそもそも難しく，同一カテゴリー内の見た目の変化が対象によって極めて大きいために，同一物体を探す特定物体認識よりも一段難しい問題となっている．例えば,「椅子」などは，様々な形状のものが存在し，何が椅子であるかを計算機に教えることは極めて難しい問題である．

一方，特定物体認識では，数百万枚のデータベースに対して 95%以上の精度で同一物体の検索が可能になっており，すでに一部で実用化が始まっている．例えば，Google が昨年 11 月に発表した，キーワードの代わりに画像を入力とする検索サービスである Google Goggles において，特定物体認識技術を用いてユーザの撮影したランドマークや有名絵画の写真認識し，自動認識された対象物の名称を使って Web 検索をするサービスが試験的に行われている．

3. 画像表現手法

本節では，まず簡単に一般物体認識の歴史について述べ，次に最新の画像表現手法について説明する．

3.1 研究の歴史

一般物体認識は，画像認識の研究が始まった 1960 年代当初より研究が行なわれていた．しかしながら，当時はそもそも計算機でカラーデジタル画像を扱うこと自体が困難で，線画を対象に線画解釈の研究が盛んに行われた．その後，1970 年代，1980 年代は，２次元的な取扱いのできる画像，例えば，航空写真や風景写真などの様な画像に対する認識システムが盛んに研究されるようになった．当時は，画像を領域分割して，各領域の形状や色，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．1980 年代には，人工知能のエキスパートシステムの手法が導入され，複雑なルールに基づく認識システムが開発された．しかしながら，認識のためのルールは人手によってすべ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，実験用画像以外の一般の画像を対象とした認識を実現することは出来なかった．

1980 年代後半になると 3 次元の実世界を対象とする認識が盛んになった．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ベーストによる物体認識の研究が盛んに行なわれた．しかしながら，認識対象の正確な形状モデルが事前に必要であるために，特定物体認識の認識しか実現できなかった．こうした認識は，カテゴリー分類でなく同一対象の検索であったために，汎化は必要でなく，学習が利用されることはなかった．

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．特に，顔画像認識では，学習を用いた方法で大きな進歩を遂げた．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法 [2] は，その代表的な方法である．また，それを一般の 3 次元物体の特定物体認識に適用するパラメトリック固有空間法 [3] 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．しかしながら，認識対象全体を特徴として利用しているので，物体の一部分が隠れたりするオクルージョンや部分的な変形に対処出来ないという問題もあった．

当時は現在のような非線型サポートベクターマシンのような高次元ベクトルに対して汎化能力を発揮する手法が存在していなかったため，主に高次元のベクトルの次元数を下げる統計的手法である主成分分析 (Pricipal Conponent Analysys, PCA)，判別分析 (Linear Discriminant Analysis, LDA)，正準相関分析 (Canonical Correlation Analysis, CCA)，それらを認識に応用した部分空間法などが主に研究され，文字認識を中心とする画像認識に応用されていた．

一方，画像認識とは異なる関連研究分野として，画像データ

ベース検索という研究分野があり，見た目が類似している画像を画像データベース中から検索する，内容に基づく画像検索（content-based image retrieval，CBIR）の研究が 1990 年代から行われるようになった．画像の色やテクスチャの分布をヒストグラム化し多次元ベクトルで画像を表現して，類似ベクトルを検索することによって，類似画像検索を実現した．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が，色やテクスチャなどの低レベル特徴のみを用いているために，見た目が類似している画像や,「暖かい」「暗い」などの主観的な形容詞に対応する画像は検索することが可能であるが，意味的に類似している画像を検索することは困難であった．そのため，CBIR の研究例は多いが，実用的に使われている例はあまり多くはない．

3.2 局所特徴量

以上述べたように画像認識の研究は研究自体は長期間行われていたものの，常に何らかの前提条件が必要で，実験画像に対してうまくいく手法はあっても，一般の人がカメラで撮影した制約のない一般的な画像に適用できる手法は存在しなかった．2000 年前後までは，一般物体認識は極めて困難な問題として考えられており，どの様にアプローチすればいいのかさえ定まっていな状況であった．そうした状況に対して，90 年代の後半から 2000 年代の前半にかけて，一般物体認識に関するブレークスルーが起った．それに関する重要な研究は,（1）局所特徴の組合せによる画像の表現,（2）局所特徴の表現法，そして,（3）局所特徴のヒストグラム表現である bag-of-features である．

まずは 1990 年代後半に，認識対象全体を用いるのではなく，認識対象の特徴的な局所パターンを多数抽出し，その組合せによって，画像検索および特定物体認識を行う方法が提案された [4]．認識に用いる特徴的な部分の抽出には，元々はステレオ 3 次元復元やパノラマ画像生成に必要な複数画像の対応点検出のために研究されてきた局所特徴抽出手法が利用された．代表的な方法としては，特徴点検出と特徴ベクトルの抽出法をセットにした SIFT（Scale Invariant Feature Transform）法 [5] がある．

SIFT 法は (1) 特徴点とその点の最適スケールの検出，(2) 特徴点の周辺パターンの輝度勾配ヒストグラムによる 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．画像中のエッジやコーナーなどの特徴的な部分が特徴点として自動的に検出され，さらにその周辺パターンに基づいてパターンのスケールと主方向が決定され，回転，スケール変化 (拡大縮小)，明るさ変化に不変な形でその周辺パターンが特徴ベクトルとして記述される．SIFT 特徴量は，回転，スケール変化，明るさ変化だけでなく，一定の範囲内のアフィン変換 (視点の移動) にも頑健であることが実験によって示されている．つまり，図 3 に示すように，1 枚の画像で特徴点が抽出されベクトルで記述されると，もう一枚の回転，縮小，明るさ変化を加えた画像でも，同じ場所から特徴点が抽出され，その点のベクトルの値もほぼ等しくなる．そうすることにより，SIFT 法で抽出した特徴ベクトルの探索のみで，異なる画像間の対応点が検出できることになる．また，SIFT 法は濃淡画像の輝度勾配を特徴量としていて，色情報を一切使っていないため，色が異なっていても濃淡の変化が似ているなら類似パターンと見なされることも特徴である．

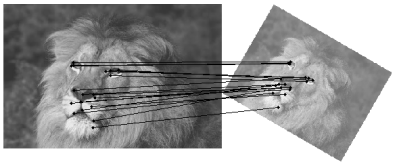

Figure 3 SIFT 特徴量を使った局所パターンのマッチングの例.

抽出する特徴点の数はパラメータによって制御可能であるが，通常は多くの対応点の候補が多数あった方がより処理が頑健になるので数百から数千個の特徴点を抽出する．そのため，多数の対応点が得られ，多少の誤対応や，部分的な隠れによる対応点の減少が起っても，ある程度の範囲内なら，物体の対応をとることが可能となる．以上が特定物体認識の基本原理である．

SIFT 法のアルゴリズム自体は実装は容易であるとは言えないが，提案者の D. Lowe 自らによるものを初め，いくつかのソフトウェアが Web 上に公開されており，手軽に利用可能となっている．なお，SIFT に関する日本語の解説としては，中部大の藤吉先生による解説 [6] が詳しい．SIFT 以外にも同様の局所特徴量は数種類提案されており，特に SURF [7] は，オープンソース画像認識ライブラリである OpenCV のバージョン 1.1 以降にライブラリ関数として取り込まれているため，手軽に利用可能である．また, [8] の HP で公開されているソフトウェアのように，SIFT 特徴に加えて，後述する bag-of-features 特徴量まで 1 つのコマンドで抽出可能な公開ソフトウェアも存在する．

3.3 Visual Words と Bag-of-features

SIFT 法に代表される局所特徴量による認識は，高精度で頑健な特定物体認識を可能としたが，1 つの画像から数百から数千のもの多数の局所特徴量を抽出すると，多数の画像に対して特徴点を高速に照合することが困難になる．そこで，1 枚の画像から多数抽出される局所特徴ベクトルをベクトル量子化し，代表ベクトルである code word に置き換えて，対応点の検索を行う手法が提案された [9](図 4)．代表ベクトルは visual word とも呼ばれ，特定物体認識を行う場合はその数はデータベースのサイズに応じて，1 万から 100 万程度の値が選ばれる．この visual words の考え方を用いると，画像から抽出された局所特徴ベクトルは単語 (visual word) に変換されるので，1 つの画像は数百から数千の単語の集合によって表現されることになる．つまり，画像は，文章や Web ページなどと同じで，単語の集合として表現されることになる．実際に visual words を提案した論文 [9] では，テキスト検索で用いられる転置インデックス法を visual words 表現された画像に適用し，テキスト検索手法を応用することで高速な特定物体認識が可能となることを示した．

Visual words の最初の論文は特定物体認識を目的としていたため，それだけでは一般物体認識への適用は不可能であった．局所特徴量および visual words を一般物体認識に応用すること可能としたのは，bag-of-features 表現 (BoF) [10] である．

文章をベクトル表現する方法として，語順を無視して単語の

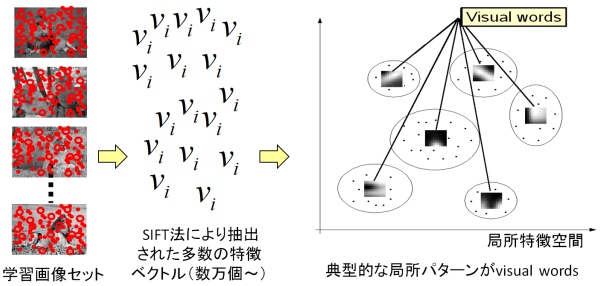

Figure 4　代表局所パターン (visual word) の求め方．認識対象の学習データセットから局所特徴特徴ベクトルを抽出し，クラスタリングで visual words を求める．

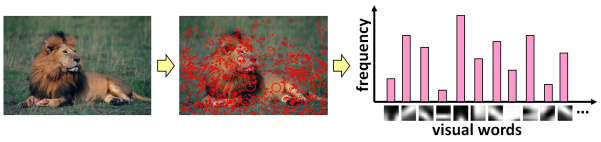

Figure 5　Bag-of-features 表現の求め方．すべての局所特徴量を visual words に対応させ，ヒストグラムを作成する．

出現頻度ベクトルで文章を表現する bag-of-words(BoW) 表現が言語処理や情報検索の分野で用いられているが，それとまったく同様に，各特徴点の画像中での位置，つまり visual words の位置を無視して，visual words を bag-of-words 化したのが，bag-of-features 表現である (図 5)．そのため，bag-of-visual-words（BoVW）と呼ばれることもある．なお，BoW と同様に TF-IDF で要素を重み付けする方法や，各要素を 1 か 0 の 2 値にしてヒストグラムを構成する方法も試みられているが，一般には，単純に visual words の出現回数をカウントしてヒストグラムを作成し，BoF ベクトルとする場合が多い．なお，BoF ベクトルは，各画像の visual words の合計頻度に差がある場合は，L1 正規化することが一般的である．

Bag-of-features は，結局，画像から抽出された局所特徴量の分布を visual words のヒストグラムで表現しているということである．ヒストグラムは，色に関しては従来より画像表現の一つとして利用されてきたが，色ヒストグラムは似た色の画像の検索には有効であったものの，色は物体のカテゴリーとは必ずしも直接結び付かないために，カテゴリー認識を目的とした一般物体認識においてはあまり有効ではなかった．それに対して，局所パターンは物体のカテゴリーと関係が深く，その分布のヒストグラムである bag-of-features は多くの一般物体認識の研究においてその有効性が示されている．さらに，注目すべきは bag-of-features の元となる局所特徴量は濃淡変化のみに注目していて bag-of-features には色に関する情報はまったく含まれていないにも関わらず，従来の色などの特徴量よりも高い精度でカテゴリー分類が可能となっており，物体のカテゴリー認識には色情報は重要ではないということが実験結果から示された形になっている．

Bag-of-features 表現はヒストグラム表現であるため，各局所パターンの位置の情報が完全に捨てられてしまっているが，逆にその潔さが表現の簡潔さにつながり，現在，一般物体認識において標準的な画像表現手法として広く使われるに至っている．

なお，一般物体認識においては同じカテゴリーに属する物体の細かな差異が吸収されることが望ましいので，visual words のサイズは，特定物体認識ほどは大きくせずに数百から数千程度である．一方，特定物体認識では，まったく同じ局所パターンだけが 1 つの visual word に割り当てられることが望ましいので，数万から百万程度のサイズが一般的である．特定物体認識においては visual words は広く用いられているが，局所パターンの分布が類似していることよりも，一致する局所パターンが一定数存在するということの方が重要であるので，分布を表現した bag-of-features ベクトルを機械学習手法で学習するようなことは特定物体認識においては行われない．

Bag-of-features の画期的な点は，bag-of-features 表現に変換された画像は文章とまったく等価に扱うことができる点である．そのため，bag-of-features が提案された直後は，競って言語処理の分野で提案された手法が画像認識に応用されるということが起った．特にカテゴリー分類においては，bag-of-words 表現が数千から数万次元もの高次元になるテキスト分類で定評のあったサポートベクターマシン（SVM）が同様に数百から数千次元になる bag-of-features ベクトルに対しても幅広く用いられている．

なお，SIFT 法などの局所特徴量抽出手法は，特徴点の検出の処理も含んでいるが，第一段階の処理の特徴点検出を用いずに，決められたピクセル毎の格子点 (グリッド) やランダムに選ばれた点を機械的に特徴点とする方法も一般物体認識においては広く用いられている (図 6)．特徴点検出手法では，空や道路の路面のような均一な領域からは特徴点が得られないが，物体カテゴリーの認識においては，テクスチャのない均一な局所特徴も重要な情報であるため，画像の内容にかかわらず機械的に特徴点の位置およびスケールを選択する方法も有効であるとされている [11], [12].

Figure 6　3 種類の特徴点検出法．円の大きさは局所特徴量のスケールを表す．グリッドの場合は同じ格子点について複数のスケールが指定されている．

Bag-of-features（BoF）はシンプルな手法であるために，その拡張が様々な面において試みられている．F. Jurie ら [13] による k-means によるクラスタリングの代りの，オンラインクラスタリングと mean-shift [14] に基づいたコードブック作成法，F. Perronnin ら [15] 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，J. Weijer ら [16] による色情報の追加，時空間局所特徴を用いた動画像への拡張などである．Dollar らは静止画像の局所特徴を時間軸方向に拡張した 3 次元の時空間局所特徴をベクトル量子化する bag-of-video-words を提案し，歩く，走るなどの人間の動作の分類に成功している [17]．Dollar らの研究によって動作認

識も物体認識と同じ BoF で手軽に扱えることが示されたため，特に，動作認識への BoF の応用は近年急速に広まっている．

4. 分類手法

前節で述べたように，bag-of-features(BoF) 表現による画像特徴ベクトル (BoF ベクトル) が抽出された後は，テキスト表現である bag-of-words ベクトル (BoW ベクトル) と等価に扱うことが可能である．BoF ベクトルは，BoW ベクトルと同様に数百〜数千次元の高次元ベクトルとなるので，高次元ベクトルの扱いが容易なサポートベクターマシンが一般には最もよく使われる．

4.1 Support Vector Machine

分類手法としては，生成確率モデルに基づく手法，サポートベクターマシンやブースティングに代表される判別モデルの基づく 2 つの方法があるが，一般に，bag-of-features を画像表現に用いる場合は，サポートベクターマシン (Support Vector Machine, SVM) を用いるのが一般的である．

SVM が画像認識で最もよく用いられる理由としては，画像認識でよく表れる高次元データに対して，SVM が次元の呪いの問題を影響をほとんど受けずに，高い汎化性能を持っていることが第一の理由である．また，問題の特性に応じたカーネル関数を利用することで，識別性能を向上させることができる点もその理由の一つである．それに加えて，品質が高く，使いやすい SVM のオープンソースソフトウェアによる実装が Web から簡単に入手できることも SVM が学習手法としてよく使われる大きな理由である．多くの一般物体認識の論文で，SVMlight [18] や LIBSVM [19] が bag-of-features 表現と組合せて利用されている．

Bag-of-features 表現による一般物体認識では，χ^2RBF カーネル，EMD カーネル [20] など様々なカーネル関数が登場している．Zhang ら [20] は，画像分類において，χ^2RBF カーネルは，最も性能が良かったカーネルの一つであると報告している．また，Zhang らは [20] で，Earth-Movers Distance(EMD) カーネルも χ^2RBF カーネルに匹敵する優れたカーネルであると報告しているが，EMD を求めるには非常に計算時間がかかるという問題がある．

χ^2RBF カーネルは，単純なユークリッド距離とは異なり，対応するベクトルの要素の和で差の 2 乗を割っているため，値の大きな一部の要素の影響が突出しないようになっており，ヒストグラム特徴である bag-of-features には適していると言われている．χ^2RBF カーネルは以下の式で表現される．

$$K(\mathbf{x},\mathbf{y}) = \exp\Big(-\gamma\chi^2(\mathbf{x},\mathbf{x})\Big)$$

$$\text{where } \chi^2(\mathbf{x},\mathbf{y}) = \sum \frac{(x_i - y_i)^2}{x_i + y_i}$$

パラメータ γ は次のように設定する．Zhang らは [20] は，χ^2 カーネルのパラメータ γ に，全ての学習画像のベクトルの組み合わせの χ^2 距離の平均の逆数を設定することによって，良い結果を報告している．距離の平均で割ることは，特徴空間の正規化を意味し，異なる特徴空間でも同じ距離尺度で比較することができる．実際，cross validation で求めた最適な γ よりも平均距離の逆数を設定した方が性能が高かったという結果もある [21].

Bag-of-features 表現は，元の特徴点の位置を完全に無視しているが，例えば，自動車なら下の方にタイヤ特有の visual words が存在し，上の方に自動車の屋根に特有の visual words が存在する可能性が高い．そこで，S. Lazebnik ら [22] は，BoF を画像全体からではなく，画像を 4 分割，16 分割して画像ピラミッドを構築して，それぞれから BoF を構築し，ピラミッドのレベルに応じて重みを付けて類似度を計算する Spatial Pyramid Kernel を提案し，それを SVM のカーネル関数として用いることで，大幅に分類性能が向上をすることを示した．このように，bag-of-features + SVM カーネルの工夫 という研究も近年多く行われるようになっている．

4.2 確率トピックモデル

一方，確率的な生成モデルとしては，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) [23]〜[25], Latent Dirichlet Allocation (LDA) [11], [26] などが一般物体認識に応用されている．Bag-of-features は統計的言語処理の bag-of-words と等価であると見なすことができるので，統計的言語処理の分野で提案された様々な確率的な手法が応用されている．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．

Latent Semantic Analysis (LSA) [27] 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 [23]．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，変分ベイズ学習もしくは MCMC をそれぞれ一般的に用いる点が異なっている [26].

J. Sivic ら [25] 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) [23] 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．予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．

他にも元々はテキスト分類のために考案されたノンパラメトリックベイズに基づく確率的手法が様々な形で一般物体認識に応用されている [28], [29].

4.3 特徴統合

Bag-of-features（BoF）による画像特徴量は画像の意味理解に有効であるが,「ひまわり」「パンダ」のような特徴的な色を持っている認識対象も存在するため，局所特徴量に基づく BoF

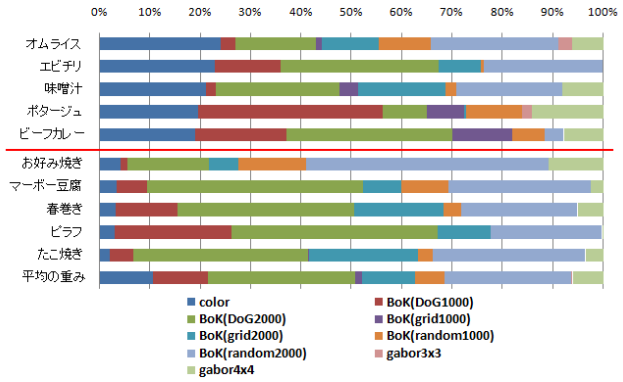

Figure 7 学習した特徴量の重み．左端の青色が部分が色の重みを示す．

Figure 8 色特徴の重みが大きかったオムライス（左上），エビチリ（右上），重みが小さかったピラフ（左下），たこ焼き（右下）の画像．

のみでなく，色やテクスチャ，エッジ特徴など様々な特徴量を対象に応じて選択的に利用することで認識精度がさらに向上できることが近年示されている．

Varma らは，BoF に加えて，テクスチャ，色，形状などの多様な画像特徴を画像から抽出し，認識対象に応じた適切な重みを Multiple Kernel Learning の手法を用いて推定し，一般物体認識を行うことを提案した [30]．Car，Face，Airplane など一般的な意味カテゴリに対応する画像を 101 種類 9,144 枚，256 種類 30,607 枚含んだ一般物体認識のアルゴリズムの性能比較のための標準的なベンチマークデータセットであるカルフォルニア工科大学の Caltech-101，Caltech-256 (注1) を用いた画像分類実験で，単一の特徴量ではそれぞれ約 65%，約 35%であった分類精度が，Multiple Kernel Learning による特徴統合によって約 90%，約 60% にそれぞれ向上した．2004 年には Caltech-101 の分類精度が 20%未満であったことから，ここ数年で急速に技術が進歩していることがうかがえる．

上東らは，同じ手法を 50 種類の食事画像の分類に応用した [21]．実験では 50 種類のマルチクラス分類を行い，平均分類率 61.34%を達成した．単体の特徴では最高でも 34.64%であったので，特徴統合によって大きく性能が向上した．図 7 に学習した特徴量の重みを色特徴の重み（青色）の上位下位 5 種類ずつの食事カテゴリを示す．図 8 に色特徴の重みが大きかったオムライス，エビチリと色特徴の重みが小さかったピラフ，たこ焼きの画像を示す．オムライス，エビチリは，それぞれ黄色と赤色が特徴的な色であるが，ピラフやたこ焼きは様々な色が含まれていて特徴的な色がないため認識には色特徴は重要ではなかった．

Figure 9 領域に単語ラベルを付けた結果．

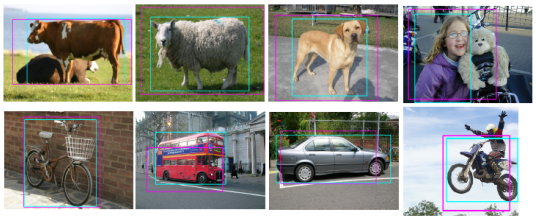

Figure 10 左上から順に cow，sheep，dog，person，bicycle，bus，car，motorbike．水色の枠が正解データ，紫の枠がスライディングウィンドウによる認識結果．

4.4 対象位置の検出

これまで説明した Caltech-101/256 や TRECVID は，画像やショット中に対象が含まれるか 1/0 分類し認識を行った．それに対して，対象が画像中や映像中のどこに含まれているかを検出する位置検出を伴う認識に関する研究も行われている．方法としては，画像を領域に分割して領域毎に認識する方法と，画像の一部分にウィンドウを設定し，それを拡大縮小してスライドさせながら，各ウィンドウに対して画像全体を分類するのと同様の方法で分類を行い，画像全体から該当物体の検出を行うスライディングウィンドウと呼ばれる方法，の 2 つが存在する．

領域に基づく方法で最も有名な方法が K. Barnard らによる word-image-translation model [31] がある．彼らは，予め画像全体対して数個のキーワードが付けられている画像データを用いて，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．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統計的機械翻訳手法を画像に適用することによって実現した（図 9）．

一方，スライディングウィンドウによる位置検出は，一般物体認識のベンチマークワークショップの PASCAL Visual Object Classes Challenge (注2) の 4 種類のタスクのうちの 1 つに detection 課題として含まれている．与えられた学習画像を用いて学習し，与えられたテスト画像のどこに物体が含まれているか検出する．20 種類の物体（person，bird，cat，cow，dog，horse，sheep，aeroplane，bicycle，boat，bus，car，motorbike，train，bottle，chair，dining table，potted plant，sofa，tv/monitor）について，個別に検出を行うが，その精度（平均適合率）は 20 種類の平均で 22%，最高でも 4 割程度に留まっている（図 10）．

(注1)：http://www.vision.caltech.edu/Image_Datasets/Caltech256/

(注2)：http://pascallin.ecs.soton.ac.uk/challenges/VOC/

Lampert ら [32] は，積分画像と，スライディングの代わりに分枝限定法を用いて，高速に一般物体の位置検出を行う Efficient Subwindow Search (ESS) を提案した．特徴点毎に事前に SVM の出力値を計算し，それを積分画像としておくことで，矩形領域の SVM の出力値を高速に計算し，探索範囲を分枝限定法で徐々に絞り込んでおくことで，単純なスライディングウィンドウに比べ計算量を 100 分の 1 程度にし，高速な部分一般物体検出を実現した．この論文は 2008 年の CVPR でベストペーパーになっている．ただし，この手法は SVM の線型性を利用しているので，カーネルは線型カーネルしか利用できないため，認識性能は十分でない場合がある．そこで，最初に ESS で位置の候補を絞って，さらに Multiple Kernel Learning で複数種類の特徴によって，高精度な一般物体検出を目指す研究も提案されている [33].

このように，現状では対象の正確な切り出しは画像全体の分類よりさらに一段難しい課題とされているために，研究が盛んになった当初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た．最近では，画像全体のカテゴリー分類に関しては最新手法によって高い分類性能が得られるようになってきているため，今後はより困難な課題である，画像からの認識対象物体の切り出しに研究の中心が移りつつある．実際，画像全体のクラス分類は，例えば，caltech-101/256 の最高精度は 2007 年の Varma ら [30] 以降大きくは向上しておらず，頭打ちの状態になりつつある一方，物体の位置検出は盛んに研究が行われている．

5. データセット

Web が普及し，容易に画像認識用のデータセットが構築可能となった．学習・テスト用のデータセットは機械学習にとっては不可欠であるので，本節では学習や評価に必要なデータセットの作成の問題について触れる．

5.1 学習データの作成

実用を目指した認識を行う場合，大規模な学習データセットが不可欠である．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．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 40,000 枚の画像に対して 20 種類の物体/シーンのアノテーションを行い，学習データの作成を行っている．また，TRECVID の 2005 年の 80 時間分のニュース映像データについて，人手によって 449 種類の学習データを作成し公開している，IBM，CMU，Columbia 大を中心とした LSCOM（Large-Scale Concept Ontology for Multimedia）(注3) [34] というプロジェクトもある．

以上は，画像全体について特定物体が含まれるかどうかラベル付けされたデータセットであるが， 大まかに領域分割された画像のそれぞれの領域にラベル付けされたデータセットを Web 上のボランティアによって構築する LabelMe プロジェクト (注4) [35] 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．

他には，画像へのアノテーションの作業をオンラインゲーム化した ESP game [36] や，オンライン上で有料で作業してもらう Amazon Mechanical Turk を用いた大規模一般画像画像データベース Imagenet.org の作成 [37] などの，ネット上の不特定多数の人に作業を行ってもらう試みもある．

5.2 Web マイニングによる認識のための知識収集

一方，一般物体認識や映像認識のための Web からの自動知識獲得の研究も試みられている．特に近年は Yahoo 画像検索 API や，Flickr API 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．

柳井 [38], [39] は Web から画像を自動収集し，それを一般物体認識のための学習データとして利用することを提案した．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この研究はその画像版ということで「Web 画像マイニング」と呼ばれている．

Web 上の知識は人手によって構築されたデータセットとは異なり，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 [40] はモデル学習時に RANSAC [41] を用いた．一方，A. Angelova ら [42] は，一般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 [43], [44] 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．Schroff らは，SVM のソフトマージンを大きく取ることによって，BoF 表現されたノイズのある Web 画像データからの学習が可能であることを示した [45].

R. Fergus らによる [24] 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．また，Vijayanarasimhan ら [46] は，Web から自動収集した正例には必ずノイズが含まれると見なして，正例集合を positive bag をみなして Multiple Instance Learning を用いて，ノイズのある画像セットからの学習を行い，一部のキーワードに関して高い性能を示した．

一方，ノイズが含まれていても大量にデータがあれば，十分

(注3)：http://www.lscom.org/

(注4)：http://labelme.csail.mit.edu/

に学習データとして利用できるという研究結果も示されている．A. Torralba らは，8,000 万枚もの大量の Web 画像を収集し，32×32 の画像に縮小して，単純な k-最近傍分類で画像分類を行った．その結果，Web から収集しただけのノイズが含まれた画像データを学習データとしてもその量が十分に多ければ，単純な手法であっても bag-of-features などの最新の手法に匹敵する一般物体認識が実現できることを示した [47].

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 から画像・映像理解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. おわりに

本稿では，一般物体認識の基本的な画像表現手法である bag-of-features について解説し，さらに BoF 表現を用いた画像分類の研究について紹介した．

Bag-of-features は，現在のところ，画像のカテゴリー分類においては最も標準的で強力な画像表現手法である．一度画像が BoF 表現になってしまえば，画像もテキストやその他の機械学習の対象データとほぼ同じように扱うことが可能であるので，今後はより多くの機械学習研究者が一般物体認識を学習アルゴリズムの実験の題材として利用することを期待したい．

References

[1] 柳井啓司: 一般物体認識の現状と今後, 情報処理学会論文誌: コンピュータビジョン・イメージメディア, Vol. 48, No. SIG16 (CVIM19), pp. 1–24 (2007).

[2] Turk, M. and Pentland, A. P.: Eigenfaces for Recognition, *Cognitive Neuroscience*, Vol. 3, No. 1, pp. 71–96 (1991).

[3] 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).

[4] Schmid, C. and Mohr, R.: Local Grayvalue Invariants for Image Retrieval, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 19, No. 5, pp. 530–535 (1997).

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

[6] 藤吉弘亘: Gradient ベースの特徴抽出 – SIFT と HOG –, 情報処理学会研究会報告: コンピュータビジョン・イメージメディア研究会, No. CVIM-160, pp. 211–224 (2007).

[7] Bay, H., Tuytelaars, T. and Van Gool, L.: SURF: Speeded up robust features, *Proc. of European Conference on Computer Vision*, pp. 404–415 (2006).

[8] van de Sande, K. E. A., Gevers, T. and Snoek, C. G. M.: Evaluating Color Descriptors for Object and Scene Recognition, *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2010).

[9] Sivic, J. and Zisserman, A.: Video Google: A Text Retrieval Approach to Object Matching in Videos, *Proc. of IEEE International Conference on Computer Vision*, pp. 1470–1477 (2003).

[10] 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. 59–74 (2004).

[11] Fei-Fei, L. and Perona, P.: A Bayesian Hierarchical Model for Learning Natural Scene Categories, *Proc. of IEEE Computer Vision and Pattern Recognition*, pp. 524–531 (2005).

[12] Nowak, E., Jurie, F., Triggs, W. and Vision, M.: Sampling strategies for bag-of-features image classification, *Proc. of European Conference on Computer Vision*, pp. IV:490–503 (2006).

[13] Jurie, F. and Triggs, B.: Creating Efficient Codebooks for Visual Recognition, *Proc. of IEEE International Conference on Computer Vision*, pp. I:604–610 (2005).

[14] Comaniciu, D. and Meer, P.: Mean Shift: A Robust Approach toward Feature Space Analysis, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 25, No. 5, pp. 603–619 (2002).

[15] Perronnin, F., Dance, C., Csurka, G. and Bressan, M.: Adapted vocabularies for generic visual categorization, *Proc. of European Conference on Computer Vision*, pp. IV:464–475 (2006).

[16] Weijer, J. v. d. and Schmid, C.: Coloring local feature extraction, *Proc. of European Conference on Computer Vision*, pp. II:334–348 (2006).

[17] Dollar, P., Rabaud, V., Cottrell, G. and Belongie, S.: Behavior Recognition via Sparse Spatio-Temporal Features, *Proc. of ICCV WS on Visual Surveillance and Performance Evaluation of Tracking and Surveillance(VS-PETS)* (2005).

[18] Joachims, T.: *SVM ^{light}*. http://svmlight.joachims.org/.

[19] Chang, C. C. and Lin, C. J.: *LIBSVM: A Library for Support Vector Machines*. http://www.csie.ntu.edu.tw/čjlin/libsvm/.

[20] 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).

[21] 上東太一, 甫足創, 柳井啓司: Multiple Kernel Learning による 50 種類の食事画像の認識, 画像の認識・理解シンポジウム (MIRU 2009) (2009).

[22] Lazebnik, S., Schmid, C. and Ponce, J.: Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories, *Proc. of IEEE Computer Vision and Pattern Recognition*, pp. 2169–2178 (2006).

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

[24] Fergus, R., Fei-Fei, L., Perona, P. and Zisserman, A.: Learn-

ing Object Categories from Google's Image Search, *Proc. of IEEE International Conference on Computer Vision*, pp. 1816–1823 (2005).

[25] Sivic, J., Russell, B. C., Efros, A. A., Zisserman, A. and Freeman, W. T.: Discovering Objects and their Localization in Images, *Proc. of IEEE International Conference on Computer Vision*, pp. 370–377 (2005).

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

[27] 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).

[28] Li, L. and Fei-Fei, L.: OPTIMOL: automatic Online Picture collecTion via Incremental MOdel Learning, *Proc. of IEEE Computer Vision and Pattern Recognition* (2007).

[29] Sivic, J., Russell, B. C., Zisserman, A., Freeman, W. T. and Efros, A. A.: Unsupervised Discovery of Visual Object Class Hierarchies, *Proc. of IEEE Computer Vision and Pattern Recognition* (2008).

[30] Varma, M. and Ray, D.: Learning the discriminative power-invariance trade-off, *Proc. of IEEE International Conference on Computer Vision*, pp. 1150–1157 (2007).

[31] 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).

[32] Lampert, C. H., Blaschko, M. B. and Hofmann, T.: Beyond Sliding Windows: Object Localization by Efficient Subwindow Search, *Proc. of IEEE Computer Vision and Pattern Recognition* (2008).

[33] Vedaldi, A., Gulshan, V., Varma, M. and Zisserman, A.: Multiple Kernels for Object Detection, *Proc. of IEEE International Conference on Computer Vision* (2009).

[34] Naphade, M., Smith, J. R., Tesic, J., Chang, S. F., Hsu, W., Kennedy, L., Hauptmann, A. and Curtis, J.: Large-Scale Concept Ontology for Multimedia, *IEEE Transaction on Multimedia*, Vol. 13, No. 3, pp. 86–91 (2006).

[35] 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).

[36] 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).

[37] Deng, J., Dong, W., Socher, R., Li, J., Li, K. and Fei-Fei, L.: ImageNet: A large-scale hierarchical image database, *Proc. of IEEE Computer Vision and Pattern Recognition* (2009).

[38] Yanai, K.: Generic Image Classification Using Visual Knowledge on the Web, *Proc. of ACM International Conference Multimedia*, pp. 67–76 (2003).

[39] 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).

[40] Fergus, R., Perona, P. and Zisserman, A.: A Visual Category Filter for Google Images, *Proc. of European Conference on Computer Vision*, pp. 242–255 (2004).

[41] 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).

[42] Angelova, A., Abu-Mostafa, Y. and Perona, P.: Pruning Training Sets for Learning of Object Categories, *Proc. of IEEE Computer Vision and Pattern Recognition*, pp. 494–501 (2005).

[43] Yanai, K. and Barnard, K.: Probabilistic Web Image Gathering, *Proc. of ACM SIGMM International Workshop on Multimedia Information Retrieval*, pp. 57–64 (2005).

[44] 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).

[45] Schroff, F., Criminisi, A. and Zisserman, A.: Harvesting Image Databases from the Web, *Proc. of IEEE International Conference on Computer Vision* (2007).

[46] Vijayanarasimhan, S. and Grauman, K.: Keywords to visual categories: Multiple-instance learning forweakly supervised object categorization, *Proc. of IEEE Computer Vision and Pattern Recognition* (2008).

[47] Torralba, A., Fergus, R. and Freeman, W. T.: 80 million tiny images: a large dataset for non-parametric object and scene recognition, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 30, No. 11, pp. 1958–1970 (2008).