

## 解 説

## 物体認識技術の進歩

Recent Progresses in Object Recognition

柳 井 啓 司\* \*電気通信大学情報工学科

Keiji Yanai\* \*Department of Computer Science, The University of Electro-Communications

## 1. は じ め に

近年、デジタル画像に対する物体認識技術が急速な発展を遂げている。2000 年前後に提案された局所パターンの表現手法および物体を局所パターンの集合として表現する手法により、「机」や「ライオン」などの一般的な物体カテゴリの認識を行う一般物体認識 [1]、および同一物体の高速な検索を行う特定物体認識 [2] の技術が飛躍的に進歩した。それ以前は、一般的な物体認識を行うことは極めて困難であると考えられていたため研究自体があまり行われず、その代わり実用的な応用が期待できる人の顔や自動車など特定の対象について認識対象ごとに専用の手法が研究されていた。ところが、近年、デジタルカメラや Web 上の写真共有サイトなどの普及により対象を限定しない一般的な実世界シーンの画像が爆発的に増大し、一般的な画像認識の手法の実現への期待が高まりつつあった。そうした状況に対して、新しい物体表現の登場による技術的なブレークスルーが起こり、現在、画像認識の研究コミュニティにおいては、対象を限定しない一般性のある物体認識技術の研究が盛んに行われるようになってきている。実際に、物体認識の研究ブームとも言える状況になっている。

そこで、本稿では、局所パターンを用いた物体表現手法を中心に、今後ロボットへの応用が大いに期待される最新の物体認識技術について解説を行う。

## 2. 一般物体認識 と 特定物体認識

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 とお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の範囲が広く、同一カテゴリ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が目的であるので、カテゴリの定義に関する問題は存在しない。

図 1 に二つの認識に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は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る。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分類が最も標準的なタスクで、複数のカテゴリ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ラベルを付与する画像ラベリング、長方

原稿受付 2010 年 3 月 1 日

キーワード: Image Recognition, Generic Object Recognition, Specific Object Recognition

\*〒182-8585 調布市調布ヶ丘 1-5-1

\*Chofu-shi, Tokyo

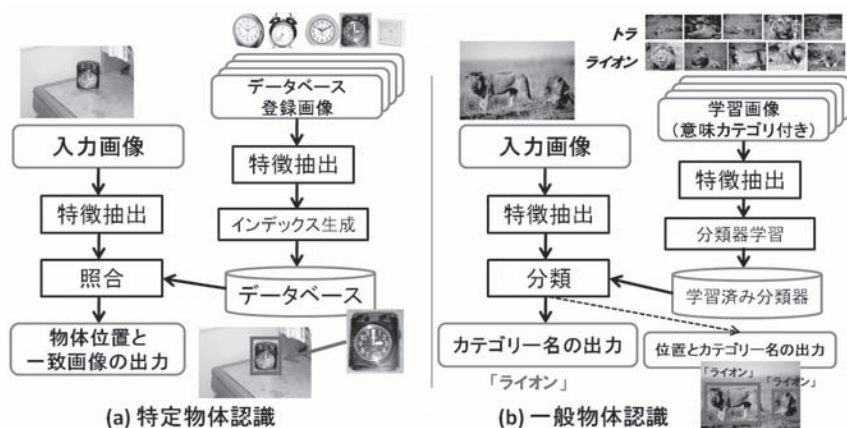


図1 特定物体認識と一般物体認識の違い。同一物体をデータベース中から検索するのが特定物体認識、意味カテゴリーを当てるのが一般物体認識。一般物体認識では、位置検出も行うと一段難しいタスクになる

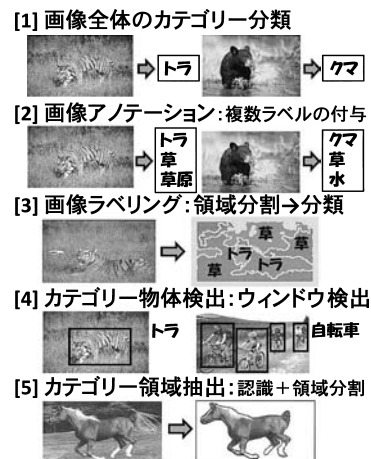


図2 一般物体認識（カテゴリー認識）の主要な5種類のタスク

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。次章では、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一つである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枚の画像に一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。

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2009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を認識し、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. 物体認識技術の発展

90年代までの画像認識研究では、対象を限定したり、画

像の撮影環境を限定したりするなど、つねに何らかの前提条件が必要で、一般の人がカメラで撮影したスナップ写真のような画像に適用できる認識手法は存在しなかった。そうした状況に対して、1990年代の後半から2000年代の前半にかけて物体認識技術に関するブレイクスルーが起こった。キーとなる研究は、(1) 局所特徴の組み合わ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。

#### 3.1 局所特徴による認識

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、その組み合わせによって、画像検索および特定物体認識を行う方法が提案された[3]。認識に用いる特徴的な部分の抽出には、元々はステレオ三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[4]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、の二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

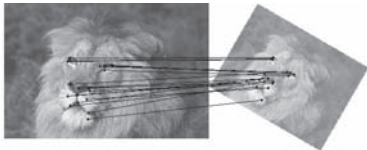


図3 SIFT 特徴量を使った局所パターンのマッチングの例

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。

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、通常は対応点の候補が多数あったほうがより処理が頑健になるので数百～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こっても、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自らによるものをはじめ、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法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[5]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[6]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。

### 3.2 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、一つの画像から数百～数千のもの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[7] (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、1万～100万程度の値が選ばれる。このvisual wordsの考え方をを用いると、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、一つの画像は数百～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[7]では、テキスト検索で用いられる転置インデックス

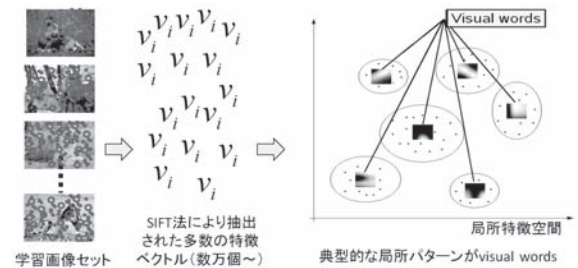


図4 代表局所パターン (visual word) の求め方。認識対象の学習データセットから局所特徴ベクトルを抽出し、クラスタリングでvisual wordsを求める



図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

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を可能としたのは、bag-of-features表現(BoF)[8]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出現頻度ベクトルで文章を表現するbag-of-words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。

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かかわらず、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～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一つの visual word に割り当てられることが望ましいので、数万～百万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ほうが重要であるので、分布を表現した bag-of-features は特定物体認識においては通常は利用されない。

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こった。特にカテゴリー分類においては、bag-of-words 表現が数千～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～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ごとの格子点 (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。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を得られないが、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。

### 3.3 Bag-of-features 法の発展

Bag-of-features はシンプルな手法であるために、visual words の生成の方法に関する工夫や、visual words の位置の情報を加味する方法、色やエッジ特徴など他の画像特徴量と統合する手法、GPU を用いた高速化など、その拡張が様々な面において試みられている。

元々の bag-of-features は画像全体を一つのベクトルで表現して、画像一枚をそのままカテゴリー分類するタスクに用いられるが、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も応用されている。主な方法としては、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、の二つが存在する。

最近では、画像全体の分類に関しては最新手法によってある程度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。

さらに静止画の認識のみならず、局所特徴を時間軸方向に拡張した三次元の時空間局所特徴量を動画から抽出し、それを bag-of-features で表現することによって、歩く、走るなどの動作分類を行う bag-of-video-words という手法も提案されている。この手法によって、動作認識も物体認識と同様の問題として手軽に扱えることが示されたため、動作認識への bag-of-features の応用は近年急速に広まっている。

## 参考文献

- [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] 黄瀬浩一: [チュートリアル] “特定物体認識”, 電子情報通信学会研究会報告, パターン認識・メディア理解研究会, no.11, 2009.
- [3] C. Schmid and R. Mohr: “Local Grayvalue Invariants for Image Retrieval,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.19, no.5, pp.530-535, 1997.
- [4] D.G. Lowe: “Distinctive Image Features from Scale-Invariant Keypoints,” International Journal of Computer Vision, vol.60, no.2, pp.91-110, 2004.
- [5] 藤吉弘巨: “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.
- [6] H. Bay, T. Tuytelaars and L. van Gool: “SURF: Speeded up robust features,” Proc. of European Conference on Computer Vision, pp.404-415, 2006.
- [7] J. Sivic and A. Zisserman: “Video Google: A Text Retrieval Approach to Object Matching in Videos,” Proc. of IEEE International Conference on Computer Vision, pp.1470-1477, 2003.
- [8] G. Csurka, C. Bray, C. Dance and L. Fan: “Visual categorization with bags of keypoints,” Proc. of ECCV Workshop on Statistical Learning in Computer Vision, pp.59-74, 2004.



柳井啓司 (Keiji Yanai)

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 年～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)。一般物体認識、Web 上のマルチメディア情報のマイニングなどに興味がある。人工知能学会、情報処理学会、電子情報通信学会、ACM、IEEE CS の会員。