3 画像識別と画像復元

基専

原田達也（東京大学）

画像識別と画像復元

画像識別

本稿では，はじめに入力画像に対してそこに映る物体やシーンのカテゴリを予測する画像識別（image classification）について述べていくことにする．近年の機械学習手法の進展や，それを支える計算機の進化，画像データセットの整備により，画像識別性能の向上は目を見張るものがある．コンピュータビジョンや人工知能分野でも注目を浴びている画像認識のコンペティション（ILSVRC）☆1は，120万枚の画像で学習し，1,000のカテゴリを予測する課題であるが，この困難な課題に対して人の識別能力と比較して同等のシステムが報告されている．ここで，画像識別を理解するために，一般的に用いられるパイプラインを図-1(a)に示し，それぞれのモジュールについて順に説明する．

1. 入力画像に対して局所的な領域の特徴を抽出する．これを局所特徴（local feature）と呼ぶ．
2. 局所特徴を識別に有利な特徴に変換する操作をコーディング（coding）と呼ぶ．後段のプーリング（pooling）を用いて画像を代表するベクトルを生成したときに，局所特徴群をモデル化した確率密度分布のパラメータとなるようなコーディング手法がよく用いられる．
3. 画像空間に配置されたコーディング後の局所特徴群を1本または少数のベクトルにまとめる操作をプーリングと呼ぶ．このプーリングには対象ベクトルの平均値を計算するものや，ベクトルの各要素の最大値を計算するものなどがある．このプーリングの結果，画像1枚を代表するベクトルが得られた場合，これを画像の特徴ベクトルと呼ぶ．
4. 画像の特徴ベクトルを人，犬，猫などのカテゴリに分類するモジュールは識別器と呼ばれている．このモジュールを経ることで画像識別が完了する．

従来の画像識別では，それぞれのモジュールを別々の問題として考えて，モジュールごとに機械学習を利用しながら設計するアプローチがとられてきた．一方，1から3のモジュールを多段に重ね，最後に識別器のモジュールを組み入れたパイプラインを考えて，初段から最終段までを一気に学習するのがディープラーニングの枠組みである．

画像復元

一方，画像識別の逆の手順をたどるのが画像復元（image reconstruction）である（図-1(b)参照）．画像復元は，さまざまなレベルから行うことが考えられる．たとえば，局所特徴レベルからの画像復元，プーリング後の画像復元，カテゴリレベルからの画像復元などである．しかしながら，画像識別の処理のレイヤが高次になればなるほど，本来画像が保持していた情報が欠落していくため，高次のレイヤからの画像復元はより困難な課題となる．

画像復元のモチベーションとしては，いくつかある．画像識別は複雑なモジュールを通じて処理されていくために，その中間的な情報を人が理解することが困難な状況にある．もし，この中間的な情報を人が理解可能な形で提示できれば，各モジュールで何が起きているのか直感的に理解できるようになり，処理に潜むエラーの発見や画像識別の性能向上の助けになる可能性がある．また，画像復元の次のステップとして画像生成，つまり高度に抽象化された状

☆1 http://www.image-net.org/challenges/LSVRC/

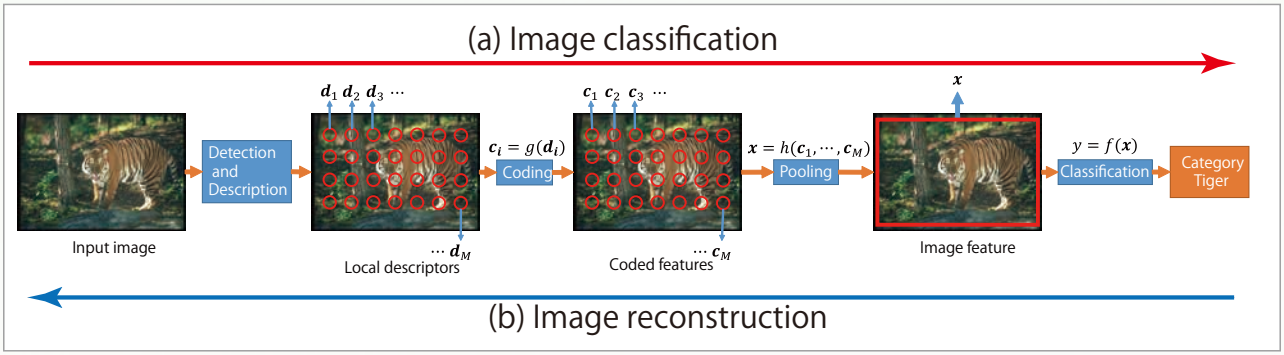

図 -1　画像識別と画像復元のパイプライン

態から新しい情報を生み出す手法につながる可能性があり，人の創造的活動を機械で実現するという側面からも興味深い話題である．以下では，画像復元に関する代表的な研究を紹介していく．

局所特徴からの画像復元

SIFT 特徴からの画像復元

局所特徴の代表として Scale Invariant Feature Transform（SIFT）特徴がある．SIFT 特徴は，画像中のコーナーのような特徴的な点を検出する検出器と，検出された点における局所領域の画像パッチ（局所の画像そのもの）内の輝度勾配ヒストグラムを表現する記述子から構成される．画像パッチが輝度勾配のヒストグラムに変換されているために，記述子から画像パッチへの逆変換は不良設定問題となる．そこで，Weinzaepfel らは，外部画像データベースから抽出された SIFT 特徴と画像パッチのペアを大量に保持していることを前提として，SIFT 特徴群から画像を復元する手法を提案している [3]．SIFT 特徴は，検出位置，輝度勾配ヒストグラム以外にも，スケール，傾き，局所領域の楕円を表現する行列，という情報を抽出している．手法の概要は以下の通りである．

1. 復元したい画像の各 SIFT 記述子をクエリとして（単にクエリと呼ぶ），最も似ている SIFT 記述子をデータベースから探し出す．
2. 最も似ている SIFT 記述子に対応する画像パッチを，クエリの保持する傾きやパッチの形状に適合するように変形する．
3. 1 と 2 の手順を復元したい画像に含まれるすべての SIFT 特徴に対して行う．
4. 変形した画像パッチ群をクエリの保持する検出位置にパッチの大きい順に張り付けていく．大きいパッチは元画像から引き伸ばされているので画像が不鮮明であるなどの問題を含む可能性があるからである．
5. パッチを張り付ける過程で，新たに張り付けるパッチと，すでに張り付けたパッチ群との重なりが生じる．ただ重ねるだけでは，復元画像の画質が良くないために，Poisson image editing を用いてパッチ間をスムーズにつないでいく．Poisson image editing とは，着目する画像領域を，その周囲の画像領域の画素値を活用して推定する手法で，画像合成では有名な手法である．
6. たとえば真っ青な空などは特徴的な点がなく，検出器で特徴点が発見されない．したがって，その領域を表現する記述子も得られないために画像復元ができない．そこでパッチが配置されない領域も 5 で用いた手法と同様の手法を用いて補完する．

図 -2(a) に元画像と，図 -2(b) に元画像から SIFT 特徴を抽出し，図 -2(c) に SIFT 特徴群から復元した画像を示す．復元画像は筆者の研究室で実装して復元したものなので，元論文 [3] の結果と若干異なる．

HOG 特徴からの画像復元

Histograms of Oriented Gradients (HOG) 特徴は物体検出によく用いられる特徴であり，SIFT 特徴と同様に局所領域の輝度勾配のヒストグラムを計算した特徴量であるが，特徴点検出を行わずに画像の

図 -2　SIFT 特徴からの画像復元．入力画像は INRIA Copydays より

グリッド上の点において密に計算する点で異なる．

ここで，**図 -3(a)** の画像から HOG 特徴を抽出し，物体検出によく用いられる Deformable Part Models（DPM）を使って人検出を行うと，図 -3(a) 右上の赤枠で囲まれた部分に人が誤って検出されてしまった．この画像からだけではなぜ人と誤検出するのか分からないし，HOG 特徴を抽出した画像（**図 -3(b)**）を見ても分からない．

そこで，Vondrick らの提案する HOG Inversion の手法 [4] を用いて，HOG 特徴を人が理解しやすい画像に復元すると，**図 -3(c)** のような画像が得られ，確かに人らしき画像が浮かび上がるために，人と誤検知してしまうことが理解できる．

この手法も，SIFT 特徴からの画像復元と同じように，局所領域の元画像と HOG 特徴のペアを外部画像データベースを利用して保持しておく．これらをそれぞれ画像基底，HOG 基底と呼ぶことにする．画像基底と HOG 基底は学習によって獲得される．復元したい HOG 特徴の局所領域を HOG 基底の重みづけ和で近似をする．得られた重みを利用して，各 HOG 基底に対応する画像基底の重みづけ和を計算することで局所領域の画像を復元する．この操作を画像全体に行うことで，HOG 特徴から画像を復元する．この手法は HOG 特徴だけではなく任意の局所特徴に適用できる．

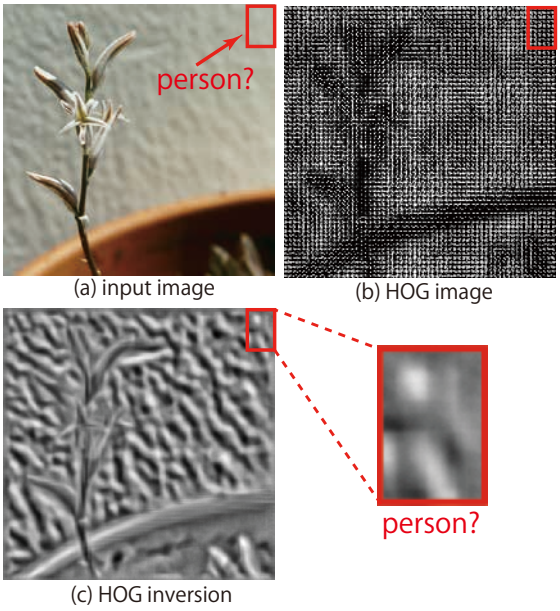

図 -3　HOG 特徴からの画像復元．入力画像は PASCAL VOC2007 より

Deep CNN からの画像復元

Deep Convolutional Neural Networks（CNN）は畳込み層とプーリング層が何層もスタックされ，最後に全結合層を組み合わせることにより実現されるネットワークである．現状の高精度な画像識別システムの多くは Deep CNN を基盤としている．Deep CNN は複雑な非線形ネットワークを多層に積み上げたシステムであり，中間でどのように処理が行われているか分からず，構造改善の方針を立てにくい．そこで，Zeiler らは Deep CNN の可視化技術（Deconvnet）を構築し，Deep CNN の性能改善につなげている [5]．Deconvnet を理解するためにまず簡単に CNN を説明する．

はじめに，畳込み層の説明をする（**図 -4** 参照）．$L-1$ 番目の層から L 番目の層の間で結合を局所に制限する．局所領域を局所受容野（local receptive field）と呼ぶ．全結合ネットワークと比較して，パ

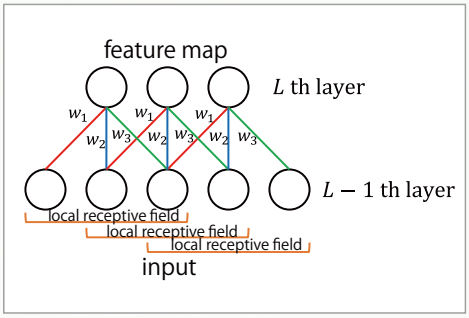

図 -4
畳込み層

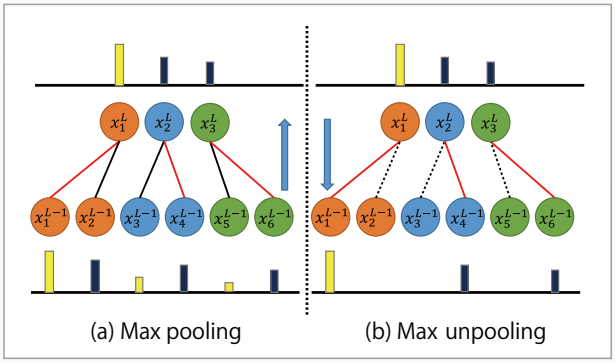

図 -5　Max pooling と Max unpooling

ラメータ数を低減させられるので汎化性能の向上が期待できる．また，画像の一部で有効な特徴抽出であれば，画像のほかの部分でも有効な特徴抽出と考えて重みの共有を行う．重みの共有の仮定によってさらにパラメータ数を減らすことができる．このようにして得られた層のことを特徴マップ（feature map）と呼ぶ．図 -4 では，局所受容野には 3 つのニューロンが存在し，それぞれ重み w_1，w_2，w_3 を用いて重みづけ和が計算される．また簡単のため図ではニューロンを直線に配置しているが，画像の場合は平面にニューロンが配置される．

重みづけ和された値は非線形活性化関数に入力される．一般に非線形活性化関数はロジスティックシグモイド関数やハイパボリックタンジェント関数などが用いられるが，これらの飽和する非線形の関数群を用いた場合，収束が遅いことが知られている．そこで $f(x)=\max(0, x)$ という ReLUs（Rectified Linear Units）を用いることで収束を高速化している．

プーリング層では最大値プーリング（max pooling）が利用される場合が多い．最大値プーリングとは上位層から接続されいる下位層のニューロン群の最大値を上位層のニューロンの値とするものである．**図 -5(a)** に模式図を示す．たとえば，x_1^L は x_1^{L-1} と x_2^{L-1} に接続されているが，x_1^{L-1} の値が高いために，x_1^L の値として x_1^{L-1} の値が採択されている．

図 -6(a) に CNN の構造を示す．CNN の画像復元はこの逆プロセスを辿ることで実現される．Deconvnet の構造を**図 -6(b)** に示す．特に問題となるのが最大値プーリングの部分である．図 -5(a) に示したように最大値プーリングでは下層から上層に伝達する際に，最大値以外のニューロンの情報が欠落してしまう．さらに，得られた最大値がどのニューロンから由来するのか，つまりニューロンの空間情報も欠落する．そこで Deconvnet では，どのニューロンから最大値が得られたのかという空間情報を保持しておき，最大値プーリングの逆の過程ではこの空間情報を利用して再構成する．**図 -5(b)** に最大値プーリングの逆変換（max unpooling）を示す．x_1^L の値は x_1^{L-1} から由来しているという情報を保持しておき，逆変換時は x_1^L の値を x_1^{L-1} に割り当てる．しかしながら，x_2^{L-1} の値は最大値プーリング時に欠落してしまっているので 0 を割り当てる．このように，Deconvnet は最大値プーリング時の空間情報を保持している場合に復元可能な手法であり，任意の上位層の特徴を復元可能な手法ではない点に注意が必要である．

ここで ILSVRC2012 でトップとなった AlexNet の情報を可視化した結果を示す．AlexNet は 5 つの畳込み層と 3 つの全結合層から構成されるネットワークである．入力は画像であり，出力は各クラスの確率である．**図 -7** は最終層から猫画像を復元した結果であり，**図 -8** は 5 番目の畳込み層から時計の画像を復元した結果である．

BoVW からの画像復元

コーディングの基本的な手法の 1 つとして Bag of Visual Words（BoVW）がある．BoVW は文章特徴である Bag of Words（BoW）のアナロジーから生まれた特徴である．BoW は単語の並び順，文

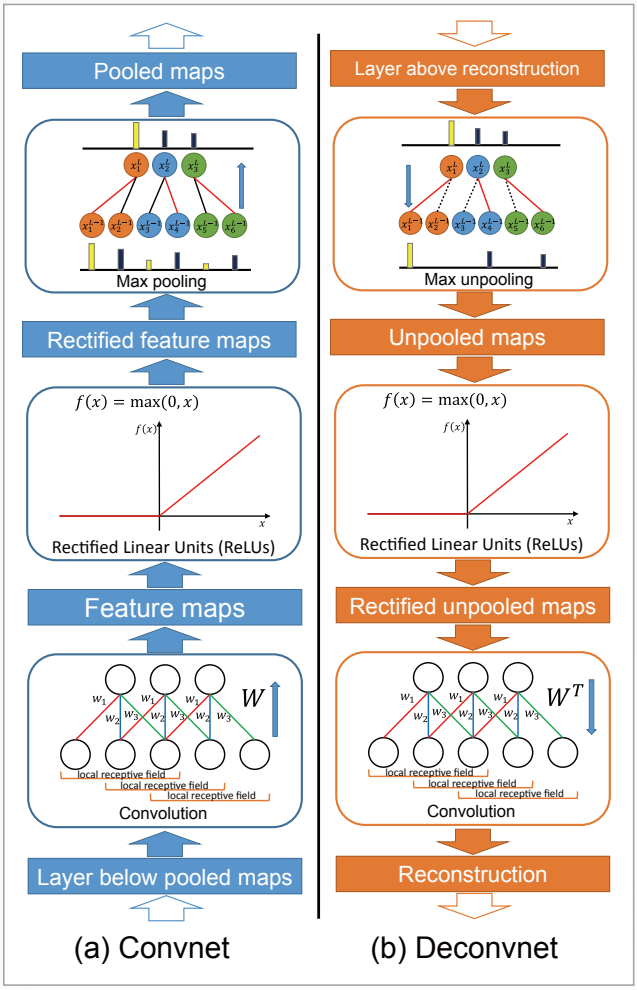

図 -6　Convnet と Deconvnet

法などを考慮しない文書特徴であり，たとえば文章中に出てきた単語のヒストグラムが利用される．BoVW は訓練集合から代表的ないくつかの局所記述子を取り上げ，画像の中に代表的な局所記述子がいくつ出現するかヒストグラムで表現したものである．これから分かる通り，BoVW からは画像にどのコードワードがどのくらい含まれるかを知ることができる．しかしながら，コードワードが画像中のどの位置から得られたかという空間情報はヒストグラムを計算する過程で失われている．その空間情報が復元されるならば，画像を再構成することが可能となる．

Deconvnet のように，画像から特徴を抽出する過程で空間情報を保持することも考えられるが，Kato らは局所記述子の空間情報を保持しなくとも画像復元が可能な手法を提案している [2]．ここでは

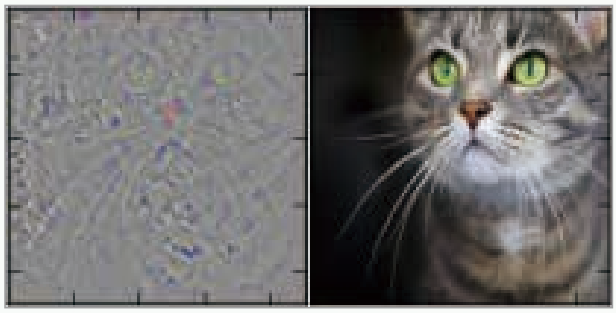

図 -7　Deconvnet による猫画像の最終層からの復元．入力画像は ILSVRC2012 より

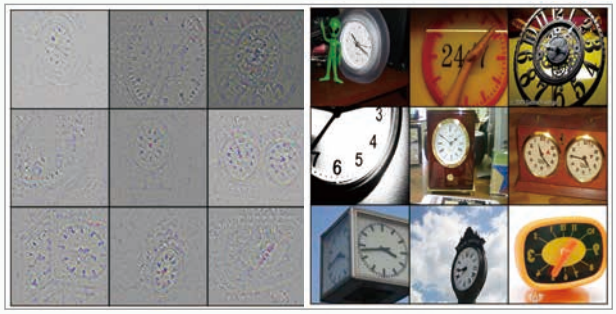

図 -8　Deconvnet による時計画像の Conv5 からの復元．入力画像は ILSVRC2012 より

BoVW Inversion と呼ぶことにする．局所記述子は画像から等間隔に抽出され，抽出に用いる画像パッチの大きさはすべて等しいものとする．BoVW から画像を再構成するために，まず各コードワードを画像上の特徴抽出点のいずれかに割り当て，次に HOG Inversion を用いて各コードワードを画像パッチへと変換する．提案手法の概略を図 -9 に示す．この手法では，コードワードの配置の仕方を評価する損失関数を構成し，その損失関数を最小化することでコードワードの配置を決定する．コードワードの配置の損出関数は，画面上で近接するコードワード対の隣り合い方の自然さ（adjacency cost）と，各コードワードの画像上の大域的位置への配置のされやすさ（global location cost）で構成されている．この損失関数を最適化する問題は NP 困難であり，局所特徴数が多いときには現実的な計算時間で厳密解を得ることはできない．そのため，遺伝的アルゴリズムと山登り法を併用した方法で近似的に最適化を行っている．

BoVW から画像を再構成した結果を図 -10 に示す．比較対象の手法として，HOG Inversion と，最

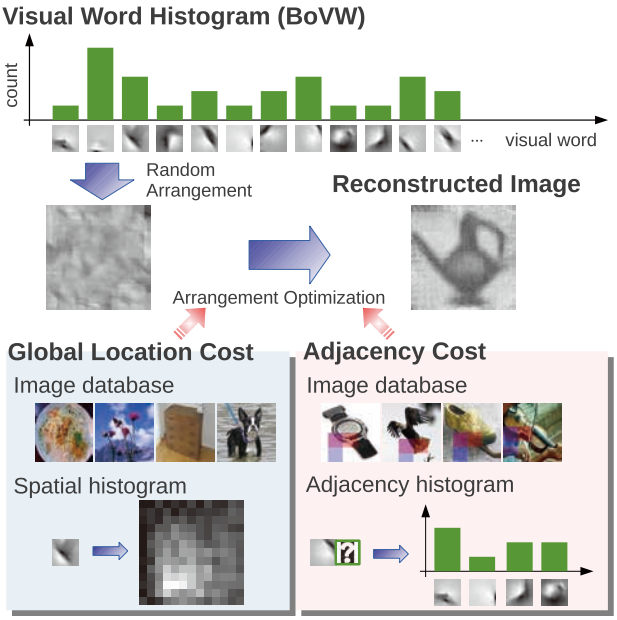

図 -9　BoVW からの画像復元の概略

近傍探索による ILSVRC2012 画像データセットからの類似画像検索を用いた．図 -10(a) に復元に用いた画像を，図 -10(b) に BoVW Inversion によって得られた画像を，図 -10(c) に HOG Inversion による画像を，図 -10(d) に検索により得られた画像を示す．BoVW Inversion により十分に理解可能な画像が得られることが読み取れる．一方で，他手法によって得られた画像から元画像の内容を推測することは困難である．

画像生成へ

画像識別の逆プロセスとしての画像復元について説明した．最近では復元のみならず新しい画像を生成する試みがいくつか発表されている．たとえば，デコーダとエンコーダの双方に Recurrent Neural Network（RNN）を組み込んだ variational auto-encoder を利用して画像生成を行っている研究がある [1]．画像復元や生成の研究は日々進展しており，人のように創造的活動を担う知的システムが出現するのもそう遠くない話かもしれない．

参考文献

1) Gregor, K., Danihelka, I., Graves, A. and Wierstra, D. : Draw : A Recurrent Neural Network for Image Generation. arXiv :1502.04623 (2015).
2) Kato, H. and Harada, T. : Image Reconstruction from bag-of-visual-words. In *CVPR* (2014).
3) Pérez, P., Weinzaepfel, P. and Jégou, H. : Reconstructing an Image from Its Local Descriptors. In *CVPR* (2011).
4) Vondrick, C., Khosla, A., Pirsiavash, H., Malisiewicz, T. and Torralba, A. : Hoggles : Visualizing Object Detection Features. In *ICCV* (2013).
5) Zeiler, M. D. and Fergus, R. : Visualizing and Understanding Convolutional Networks. In *ECCV* (2014).

(2015 年 4 月 8 日受付)

謝辞　図の作成に筆者の研究室元所属の加藤大晴氏，現所属の真野哲彰氏の協力をいただきました．

原田達也（正会員）　harada@mi.t.u-tokyo.ac.jp

2001 年東京大学大学院工学系研究科機械工学博士課程修了．2013 年同大学院情報理工学系研究科教授．現在に至る．実世界知能システム，画像認識，コンテンツ自動生成などの研究に従事．

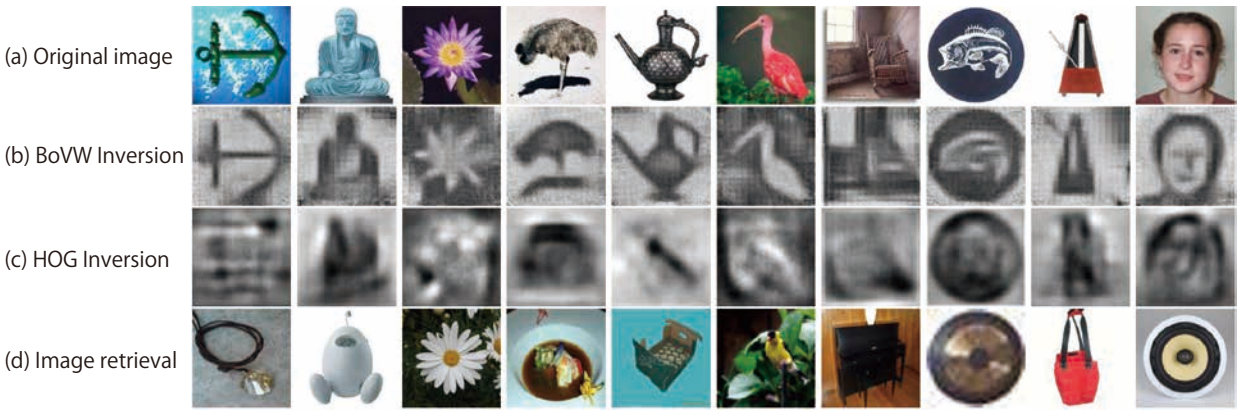

図 -10　BoVW からの画像復元の例．復元する画像は Caltech101 から．画像検索の画像は ILSVRC2012 から