局所特徴量と事例検索に基づく高速特定物体認識

岩村　雅一・黄瀬　浩一

Fast Specific Object Recognition Based on Local Feature and Example Search

Masakazu IWAMURA and Koichi KISE

Object recognition is a task to recognize an object contained in an image. If the task is intended to identify an instance rather than a category, it is called “specific object recognition.” In the current article, we introduce the specific object recognition and its applications. We also introduce a simple method using local features and efficient example search to realize the specific object recognition. The simple method can process a large-scale specific object recognition problem quickly; it realizes real-time recognition. For example, in the case that a database of 100,000 images was used, it took 100 ms per query image (excluding time required for feature extraction) to achieve 98% accuracy.

Key words: specific object recognition, local feature, approximate nearest neighbor search, voting, scalability

物体認識とは，画像中の物体が何かを言い当てる処理である．例えば，犬の写真が与えられたとき，それが「犬」である，あるいは（その犬の名前である）「ポチ」であるなどと答えることが目的である．答えが1つでないのは，物は一般に複数の概念をもつからである．概念は図1に示すように階層構造を成している．そのため，どのレベルの情報が必要かによって望ましい答えが異なる．前述の例では，「犬」という答えは動物の種類を表すのでカテゴリレベルであり，「ポチ」は個々の犬を表すインスタンスレベルである．その間には，カテゴリレベルより細かく犬種で分類したサブカテゴリレベルなどが存在する．現在のところ，異なる表現レベルの物体認識を1つの手法で実現することは難しい．それゆえ，それぞれの答えを導き出す問題は区別されており，それぞれ異なる実現方法が提案されている．「犬」と答える問題は一般物体認識（generic object recognition）[1]，「ポチ」と答えるのは特定物体認識（specific object recognition）[2–5] とよばれており，それらの中間である「チワワ」に相当するのは fine-grained object recognition とよばれている．

本稿では特定物体認識について，応用例と簡単な実現方法を紹介する．応用を考えると，特定物体認識には数千，数万，場合によっては数百万以上の物体の区別が求められる．さらに認識は高速でなければならない．このような大規模化，高速化の要請を実現するための工夫とその実現例も紹介する．

1.　特定物体認識の応用

特定物体認識は個々の物体を認識して，物体に割り当てられた ID に変換する処理とみることができる．これによってどのようなことができるのかをみてみよう．まず，物体を撮影することにより，その物体に関連付けられた情報やサービスにアクセスすることができる．図2は，ユーザーが携帯電話で印刷されたカタログを撮影することで通販サイトにアクセスしようとする例である．図2（a）のようにカタログにバーコードが印刷されていれば，バーコードを読み取る従来の技術を使って通販サイトにアクセスすることができるが，バーコードはしばしば物体の外観を損ねてしまう．しかし，図2（b）のように商品の写真（物体）そのものがバーコードの役割を果たすことができれば，その心配はない．このように，特定物体認識を用いれば，物体をバーコードの代わりに使用して特定の Web サイトへのアクセスが実現できる．Web サイトに限らず，

大阪府立大学大学院工学研究科（〒599–8531 堺市中区学園町 1–1）　E-mail: masa@cs.osakafu-u.ac.jp

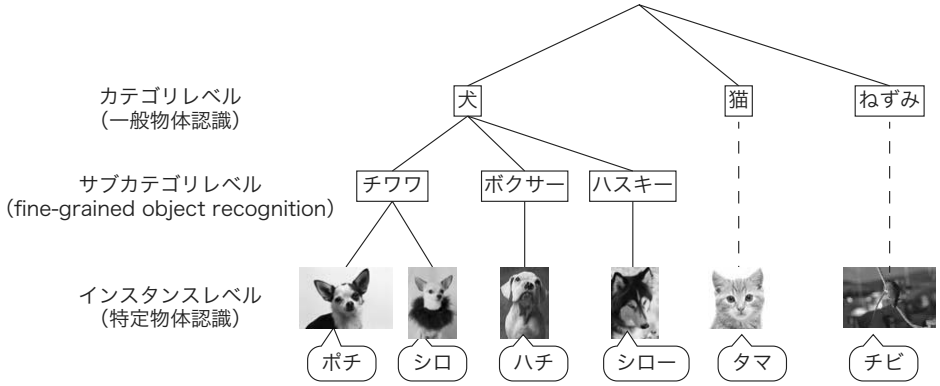

図1　物体認識の分類.

(a) バーコードの利用　　(b) 物体認識の利用

図2　物体認識の応用例：携帯電話を利用した通販サイトへの誘導.

音を鳴らす，ビデオを再生する，拡張現実（現実には存在しない物をあたかもそこにあるかのように知覚させる技術）を表示するということも可能である．見方を変えると，このことはWebの世界の「リンクをクリックする」という操作を実世界の「物体」に拡張したことに相当する．このような概念は，Internet of Things（モノのインターネット）とよばれる．

ここまでは物体を利用した情報の取り出しについてみてきたが，情報の記録も可能である．例えば，最近ユーザーによる商品のカスタマーレビューをよく見かけるが，これをWebサイトを経由せずに，商品を直接撮影することで実現することもできる．これは商品に限らず，レストランの評価などにも適用できる．前述の通販の例（図2）でも，バーコードを使うのであれば事前にカタログに印刷しておく必要があるが，物体認識であればその必要はない．レストランのレビューを投稿する際にも，自前のバーコードを店頭に貼り付けると店主に怒られてしまうが，撮影するだけであればその心配もないだろう．

別の応用としては，写真の自動索引付けが挙げられる[6,7]．観光地で撮った写真に著名な建物が写っている場合に，建物の名前を自動的に索引として付与することができる．このようなことを手軽に実現するスマートフォン・アプリケーションがすでに利用可能である．Google Goggles[1]は，ランドマーク，本の表紙，絵画，ワインラベル，ロゴなどの特定物体を認識し，それに対する検索結果を表示するアプリケーションである．このアプリケーションの興味深いところは，Web上で検索可能な画像はほぼすべてが検索できることである．すなわち，Web上の検索エンジンで適当な画像を検索して画面に表示し，それをスマートフォンで撮影して検索すると，数秒で同じ画像が検索結果として出てくる．Wikipediaによると Google では2010年までに100億枚の画像が検索可能になっている[2]ことか

[1] http://www.google.com/mobile/goggles/
[2] http://en.wikipedia.org/wiki/Google_Images

図3　画像の変動.

ら，現在はさらに多くの画像が検索対象になっていると思われる．Google は並外れた並列計算機を使うにせよ，大量の画像を短時間で検索するためには工夫が必要である．

2.　特定物体認識の実現とその大規模化，高速化

特定物体認識を実現するための最大の課題は，物体の見えの変化への対応である．同じ物体を撮影しても完全に同じ画像が得られるわけではない．その変化を吸収する方法が必要である．

もうひとつ，避けては通れないのが大規模化と高速化である．前述の応用を考えてみると，10 や 20 の物体を区別できただけでは不十分である．とはいえ，数千，数万，それ以上のオーダーの物体の認識に要する時間が，物体数に比例したのではとても使い物にならない．そこで，大規模な特定物体認識を高速に実現することが必要になる．

以下では，これらの解決策を述べる．

2.1　見えの変化に対する解決策：局所特徴量

物体の見えの変化とは，図3に示すように，同じ物の写真を撮ったとしても，必ずしも同じ画像にならないことである．これを引き起こす要因として，幾何学的変動（平行移動，拡大縮小，回転，射影歪みなど），照明変動，隠れが挙げられる．これらの問題は，局所特徴量とよばれる特徴を用いることで，ある程度解決される．局所特徴量は，その名の通り局所から特徴を抽出する方法であり，Scale-Invariant Feature Transform（SIFT）が代表的な手法として知られている[8–10]．以下では，SIFT が前述の変動に対する頑健性をどのように実現しているのかについてみていく．

2.1.1　幾何学的変動への対処―同じ領域の選択

変動を受けても同じ特徴量を抽出するために必要なのは，変動に依らずいつも同じ領域を特徴抽出に用いることである．以後，特徴量を抽出する領域を，特徴領域とよぶことにする．図4に示すように，SIFT はある画像（a）とその画像を相似変換（並進，回転，拡大縮小）した画像（b）で同じ特徴領域（図中の赤い矩形領域）を抽出可能である．相似変換のうち，並進，回転については，これらの変動に不変な計算によって同じ特徴領域を算出する．残る拡大縮小については，元画像を縮小して得られる複数の画像に対して上記の処理を行うことで特徴領域を算出する．学習時と認識時で画像の拡大率が違ったとしても，複数の特徴領域から複数の特徴量を抽出しておくことによって，特徴量の一部は一致することが期待できる．

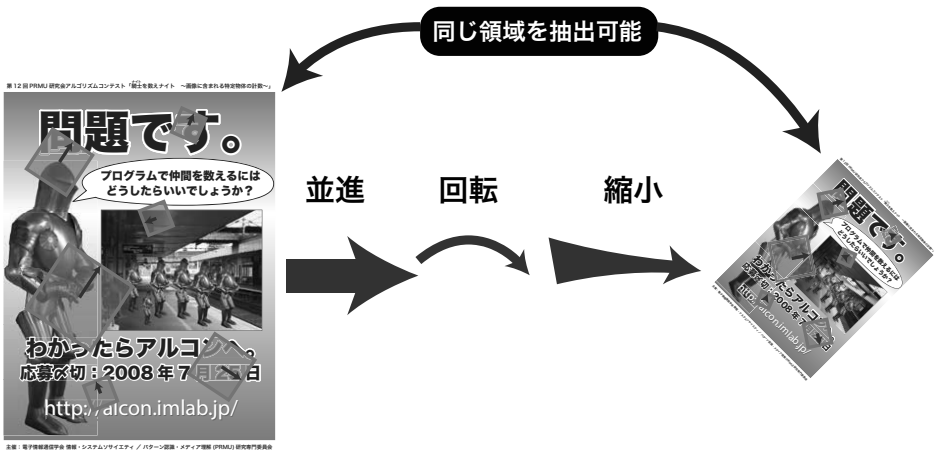

図4　並進，回転，拡大縮小に依らず同じ領域を抽出可能.

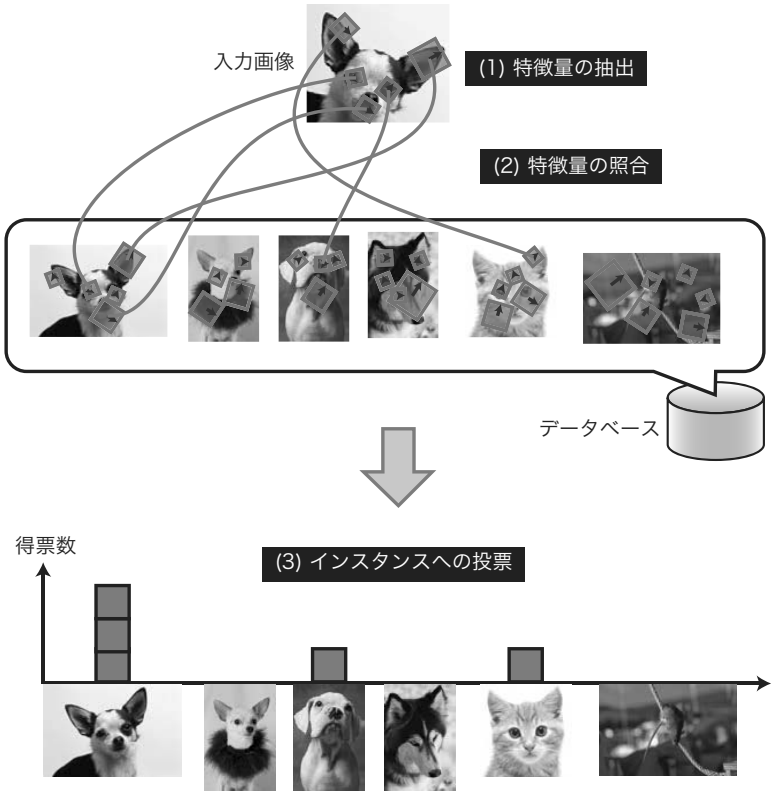

図5　局所特徴量の照合と投票による認識.

2.1.2　照明変動への対処―微分特徴の使用

撮影時の照明条件が変化すると，得られる画像が変化する．図3（c）の例では画像上部が暗く，中央部が明くなっており，図3（a）のように照明が一様ではない．しかし，画像の一部（局所）だけを見れば，照明条件はほぼ一様とみなすことができる．したがって，画像の局所から特徴量を抽出することは照明変動への頑健性にも寄与している．さらに，特徴量の計算においても照明変動に頑健な処理が行われている．隣接する画素は同じ照明条件とみなすことができるので，それらの差分は照明条件に依らない．SIFTは隣接する画素の差分（微分）に基づいて特徴量を計算するので，得られる特徴量は照明変動に頑健といえる[3]．

2.1.3　隠れへの対処―一部の特徴だけでも認識

局所特徴を用いる利点のひとつとして，隠れに対する頑健性がある．隠れがあると，一部の特徴が検出されないことや，特徴量の数値が大きく変わることが考えられる．しかし，そのような場合でも，隠れの影響を受けていない領域からは隠れのないときと同じ特徴量を抽出することができる．したがって，残った特徴のみを用いて認識することができれば，隠れに頑健な認識が実現できる．このような認識方法の中で最も単純と思われる方法は，以下のようなものである．図5に示すように，まず（1）入力画像から局所特徴量を抽出し，（2）あらかじめデータベースに登録された局所特徴量と照合し，距離が最も近い局所特徴量を照合結果とする．これにより，入力画像の各局所特徴量はデータベース中の局所特徴量の1つとそれが抽出された画像（インスタンス）と対応付く．そして，（3）局所特徴量が対応付いたインスタンスに投票する．最終的に最も多くの得票を得たインスタンスを認識結果とする．このような単純な方法であっても，局所特徴量の識別性，再現性のおかげで，時間さえかければ特定物体認識が実現できる．

2.2　大規模化と高速化に対する解決策：近似最近傍探索と投票処理の組み合わせ

前述の単純な認識方法を実装するとき，すべての局所特徴量の組み合わせで距離を計算するのが最も簡単な方法である．しかし，この方法ではデータベース中の局所特徴量の数に比例した計算時間が必要になる．計算時間の大部分は局所特徴量の照合のための距離計算で占められるため，距離計算に要する計算時間を削減することが高速化を考える上で有効な手段となる．

[3] 特徴領域の回転角の決定においても同様に微分特徴を利用している．

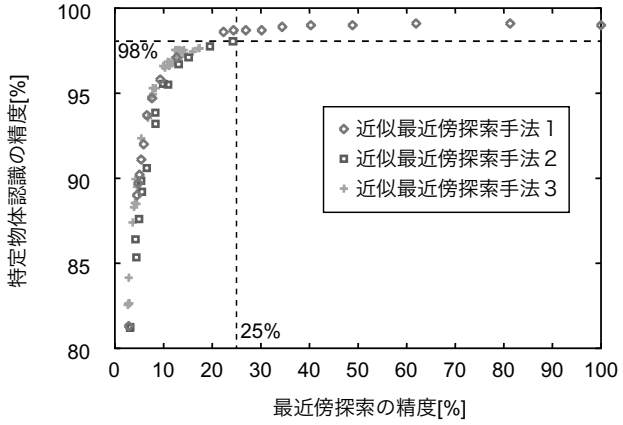

図 6　最近傍探索と物体認識の精度の関係.

局所特徴量の照合はデータベース中の最も近い局所特徴量を探す問題なので，最近傍探索問題とよばれる．物体認識の高速化において現在主流になっているのが，最近傍探索問題に近似を導入する方法である．これは最近傍探索問題において最近傍点が必ず正しく求まることを諦める代わりに，大幅な高速化を狙うものである．このように最近傍探索問題に近似を導入した問題を，近似最近傍探索問題とよぶ．最近傍探索の精度（正しく最近傍点が求まる割合）と計算時間にはトレードオフの関係があるため，時間をかければ精度は増すが，逆に精度が低くてもよいのであれば，計算時間を大きく減らすことができる．

実は，ここで導入した近似最近傍探索は，前節で導入した投票処理と相性が良い．このことを確認するために，特定物体認識の精度とそのときの最近傍探索の精度を表した図 6 をみてみよう．データベースに登録した物体数は 10,000 個であり，3 種類の近似最近傍探索手法を試した結果である．このグラフを見ると，まず 3 つの近似最近傍探索がほぼ同じ曲線上に乗っていることがわかる．そして，物体認識の精度を高くするためには，必ずしも高い最近傍探索の精度は必要ないことがわかる．例えば，4 回に 1 回の割合で最近傍探索が成功すれば，物体認識の精度を 98% にすることができる．最近傍探索の精度が低くてもよいため，計算時間を大幅に削減することができる．

図 6 でみた最近傍探索の精度が低くても物体認識の精度が高いという現象は，投票処理に因っている．このことは初歩的な確率計算で確認できる．データベースに登録された物体数を N として，もし投票処理が完全にランダムに行われるとすると，ある物体に 1 票投じられる確率は $1/N$ である．入力画像から局所特徴量が n 個抽出され，そのうち k 個が特定の物体 A に投票される事象を考えると，その事象が起こる確率は二項分布で導かれる．n 個の局所特徴量のうち k 個が物体 A に投票され，残り $n-k$ 個がそれ以外の物体に投票される確率は，${}_nC_k(1/N)^k(1-1/N)^{n-k}$ で表される．具体的な数字を入れてみよう．例えば，物体数が 1 万で，4 回に 1 回しか最近傍点が求まらない場合として，$N=10000, n=8, k=2$ とすると，2.8×10^{-7} という非常に小さな確率が得られる．このようなことは偶然ではほとんど起こらないため，もしこのようなことが起きたならば，認識結果は信頼に値するといえる．このように，最近傍探索の高速化と引き換えに最近傍探索の精度は失われるが，投票処理と組み合わせることによって，物体認識の精度を高く保つことが可能になる．

図 7　データベースに登録した画像の例.

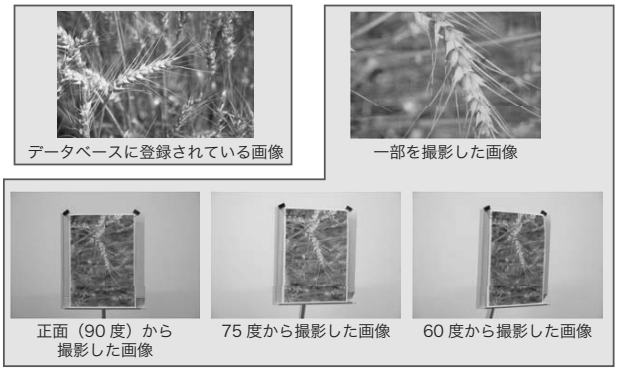

図 8　検索質問画像の例.

3.　認識実験の例

ここまで紹介してきた認識方法で認識実験を行った例を紹介する[11].

データベース用として，Web の画像検索や画像共有サイトなどから，ポスター，本の表紙，自然写真，人物の写真などの画像を合計 10 万枚収集した．このうち，500 枚，1,000 枚，5,000 枚，10,000 枚，50,000 枚，100,000 枚を抜き出し，6 つのデータセットを作成した．大きいデータセットは小さいデータセットを包含している．一例を図 7 に示す．検索質問画像として，最も小さいデータセットの 500

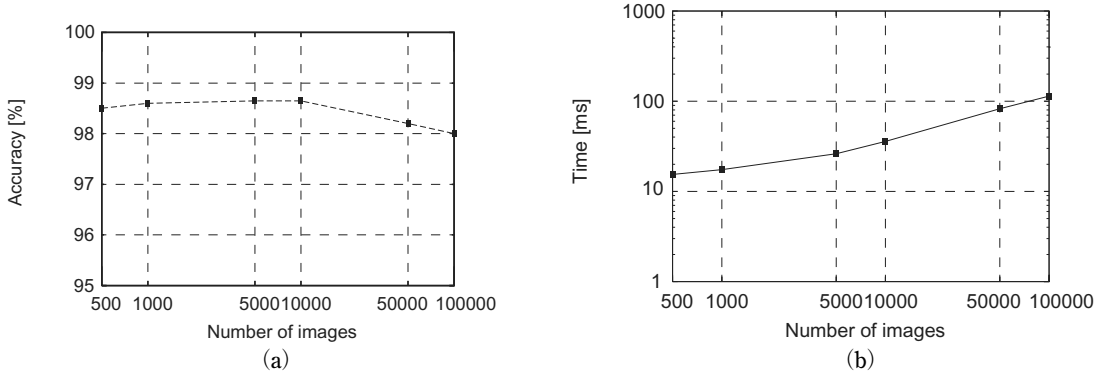

図 9　データベースの大きさを変えたときの認識率と計算時間の変化．（a）認識率の変化，（b）計算時間の変化．

枚を印刷して，正面（90 度），75 度，60 度ならびに正面向きで画像の一部のみを撮影した．500 枚×4 パターンで，合計 2,000 枚を用意した．図 8 に一例を示す．

局所特徴量として，SIFT の改良版である PCA-SIFT[12] を用いる．これは，特徴量の計算方法を改良したものであり，特徴領域の決定方法は SIFT と同じである．使用した計算機は，CPU が AMD Opteron 2.8 GHz，メモリー容量が 32 GB のものである．

データベースの大きさを変えたときの認識率と計算時間の変化を図 9 に示す．認識率はデータベースが大きくなると徐々に低下しているが，今回用いた品質の画像であれば，データベースに 10 万枚を登録した場合でも 98% 程度の認識率を達成することができた．それに要する処理時間はデータベースが大きくなれば徐々に増加するが．データベースに 10 万枚の場合，局所特徴量の抽出に要する時間を除いて 100 ms 程度であった．局所特徴量の抽出に要する時間は画像の大きさに大きく依存するが，標準的な PC を用いて数百ミリ秒から数秒程度かかる．最近は SIFT より高速で省メモリーな局所特徴量が提案されており[13]，iPad などの携帯端末で実時間で動作することもできる．

本稿では，特定物体認識の応用と大規模化，高速化の基本的な考え方を概観した．特定物体認識のより詳しい解説は文献 2–5）などで得られる．局所特徴量の最近の発展については文献 13）がわかりやすい．実装に興味がある読者は，今となっては少し古くなってしまった感があるが，文献 14）に具体的な実装方法が記載してあるので，参考にされたい．

本研究の一部は日本学術振興会科学研究費補助金基盤研究（B）（22300062）と基盤研究（A）（25240028）の成果を反映している．

文　　献

1）柳井啓司：“Bag-of-Features に基づく物体認識 (2) ―一般物体認識―”，CVIM チュートリアルシリーズコンピュータビジョン最先端ガイド，第 3 巻（アドコム・メディア，2010）pp. 85–117.
2）黄瀬浩一：“Bag-of-Features に基づく物体認識 (1) ―特定物体認識―”，CVIM チュートリアルシリーズコンピュータビジョン最先端ガイド，第 3 巻（アドコム・メディア，2010）pp. 63–84.
3）黄瀬浩一：“講座：第 2 回 マルチメディア検索の最先端 大規模静止画像 db の検索”，映像情報メディア学会誌，**64**（2010）192–197.
4）黄瀬浩一：“局所特徴量を用いた画像照合による特定物体認識”，人工知能学会誌，**25**（2010）769–776.
5）内田祐介，酒澤茂之：“大規模特定物体認識の最新動向”，電子情報通信学会誌，**96**（2013）207–213.
6）S. Gammeter, L. Bossard, T. Quack and L. V. Gool: “I know what you did last summer: Object-level auto-annotation of holiday snaps,” *Proc. International Conference on Computer Vision (ICCV)*（2009）p. 8.
7）Y.-T. Zheng, M. Zhao, Y. Song, H. Adam, U. Buddemeier, A. Bissacco, F. Brucher, T.-S. Chua and H. Neven: “Tour the world: Building a web-scale landmark recognition engine,” *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*（2009）pp. 1085–1092.
8）D. G. Lowe: “Distinctive image features from scale-invariant keypoints,” Int. J. Comput. Vis., **60**（2004）91–110.
9）藤吉弘亘，山下隆義：“物体認識のための画像局所特徴量”，CVIM チュートリアルシリーズコンピュータビジョン最先端ガイド，第 2 巻（アドコム・メディア，2010）pp. 1–59.
10）藤吉弘亘：“画像局所特徴量 sift と最近のアプローチ”，人工知能学会誌，**25**（2010）753–760.
11）野口和人，黄瀬浩一，岩村雅一：“近似最近傍探索の多段階化による高速特定物認識”，電子情報通信学会論文誌 D，**J92-D**（2009）2238–2248.
12）Y. Ke and R. Sukthankar: “Pca-sift: A more distinctive representation for local image descriptors,” *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR) vol. 2*（2004）pp. 506–513.
13）藤吉弘亘，安倍　満：“局所勾配特徴抽出技術― sift 以降のアプローチ―”，精密工学会誌，**77**（2011）1109–1116.
14）黄瀬浩一，岩村雅一：“3 日で作る高速特定物体認識システム”，情報処理，**49**（2008）1082–1089.

（2013 年 5 月 7 日受理）