Vol.2016-MUS-111 No.9
2016/5/21

画像情報処理における特徴表現

藤吉 弘亘[1,a)]

概要：音と画像の処理・認識には類似した課題が存在し，類似した技術が使われている．本稿では画像認識分野において，特徴抽出や識別手法のアルゴリズムとその変遷について紹介し，両者の類似性と差異について議論したい．

1. はじめに

本稿では画像認識分野の進展について振り返り，今後の可能性について述べる．物体認識の処理過程を，特徴点検出・記述，特徴抽出，識別器（統計的学習手法）の三つに分別し，その技術動向を紹介する．図 1(a) に画像認識における各処理過程を縦軸とした際の各処理における技術とその遷移を示す．図 1(b) は，各技術とその応用として製品化例を示す．

2. 特徴点検出・記述

画像のスケール変化や回転に不変な特徴量を抽出する Scale-Invariant Feature Transform (SIFT) [13] は 1999 年に提案され，特定物体認識だけではなく画像合成や画像分類など多くのアプリケーションに利用されている．SIFT の処理過程は，キーポイント検出と特徴量記述の二段階からなる．キーポイント検出処理では，Difference-of-Gaussian (DoG) 処理によりキーポイントのスケールと位置を検出する．特徴量記述では，スケール内の勾配情報からオリエンテーションを求め，キーポイント周辺領域（パッチ）をオリエンテーション方向に回転させて特徴量を記述することで，回転に対して不変な特徴量を抽出する．SIFT では，キーポイント検出処理における DoG 画像の生成や，特徴量記述処理における勾配ヒストグラム算出の計算コストが高いという問題がある．この問題を解決する高速化の手法として，2006 年に SURF [2] が提案された．SURF では，各処理において積分画像を利用した Box フィルタを用いることで．SIFT と比較して約 10 倍の高速化を実現した．一方，ハードウェアによる高速化として，SIFT アルゴリズムの GPU (Graphic Processor Unit) 上での実装が取り組まれた．

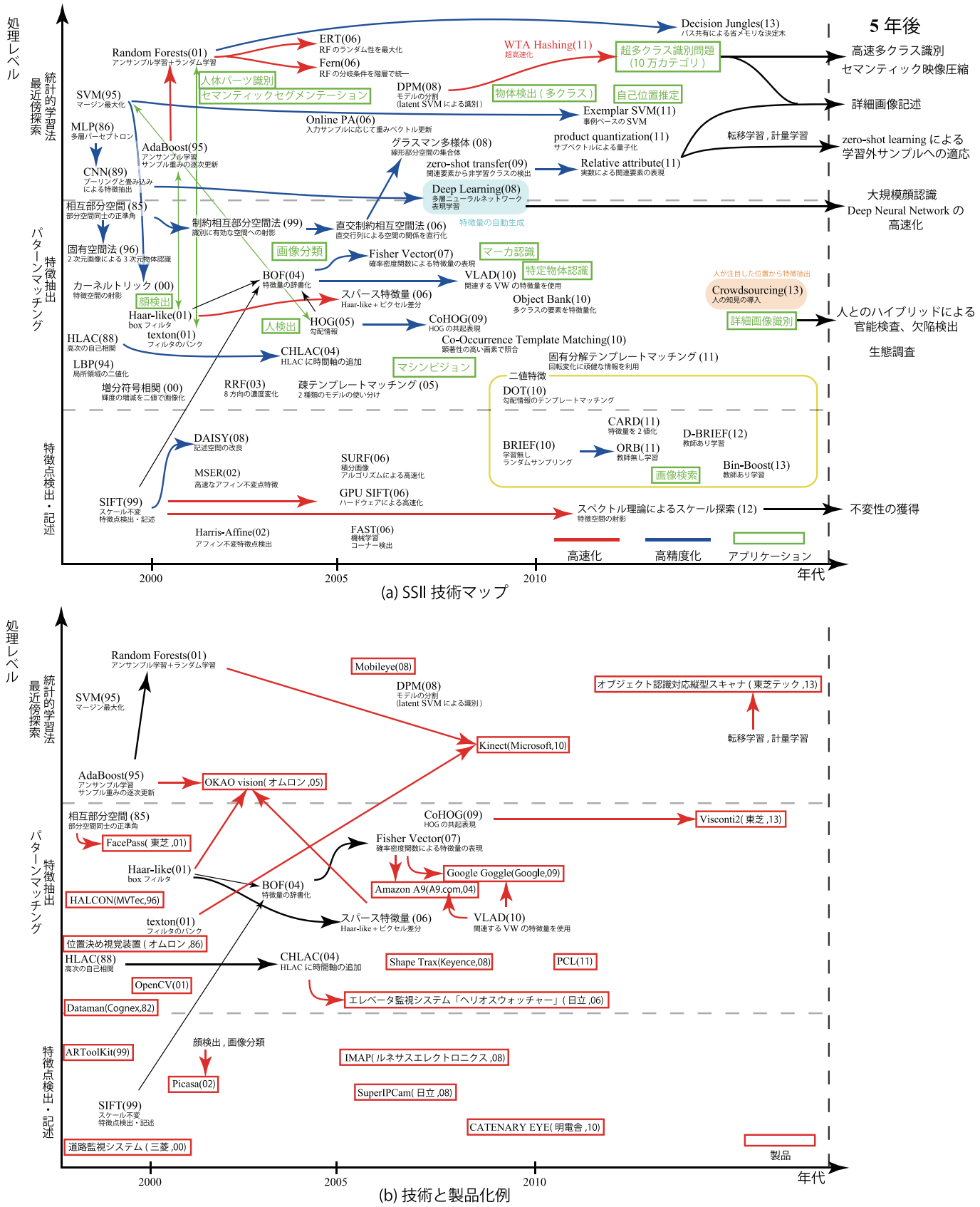
2010 年以降では，省メモリ化と高速化を同時に解決するアプローチとして，特徴量をバイナリコードで記述する手法 BRIEF [4] が提案され，局所特徴量の新しい展開を迎えた．バイナリコード表現は実数ベクトル表現に比べてコンパクトであるため，メモリの消費量を大幅に減らせるというメリットがある．さらに，バイナリコード間の類似度をハミング距離で測れるというメリットもある．ハミング距離は，2 つのバイナリコードの XOR を計算し，1 が立っているビットの数を数えるだけで得られるため，極めて高速に計算可能である．また，Streaming SIMD Extensions (SSE) と呼ばれる CPU の拡張命令セットに実装されているビットカウントを用いると，さらなる高速化を実現できる．2011 年に提案された ORB [17]， CARD [1] は，機械学習を用いてより短いビット長でより性能の高いバイナリ型特徴量を算出するアプローチである．これまでは，機械学習の利用は認識過程における後段処理として利用されていたが，昨今では前段の特徴点検出・記述に利用され始めているのは興味深い．また，ここで紹介したバイナリコード型局所特徴量の一部は既に OpenCV に実装されており，多くのアプリケーションにおける評価が容易であるため，実用化を加速させる要因となっている．このように特徴点検出・記述における技術は，認識性能の向上を目的としただけでなく，計算処理の高速化に着目したアプローチに遷移しつつあり，実利用が始まっている．

3. 特徴抽出

高精度な物体検出を実現するには，検出対象である物体に共通する情報を捉える特徴量を設計することが重要となる．顔検出に用いられる Haar-like 特徴量 [18] は 2001 年に提案され，数年後にデジタルカメラ等に利用されるようになった．人物の顔は，共通して鼻筋や頬は明るく，目や

1 中部大学
〒 487–8501 愛知県春日井市松本町 1200 番地
a) hf@cs.chubu.ac.jp

ⓒ 2016 Information Processing Society of Japan



(a) SSII 技術マップ

(b) 技術と製品化例

図 1 画像認識分野における技術の変遷. *1

口は暗い. このような人間の顔に共通するような明暗を捉えるために, 近接する 2 つの矩形フィルタ内の輝度和の差分を Haar-like 特徴量とした. また, より簡単に特徴量を表現するピクセル差分特徴量 [19] や, 矩形フィルタの

*1 本マップは SSII2014 にて作成されたものである. http://www.ssii.jp/14/special_map.html

関係性を捉えることで高精度化を果たした Joint Haar-like 特徴量 [15] やスパース特徴量 [11] が提案された．

顔検出技術が進展し実利用化が始まると共に，2005 年頃から物体検出における研究対象が顔からより難しいとされていた人に遷移した．人検出では，着ている衣服が個々に異なるため，人に共通する明暗の関係を捉えることができない．そのため，人に共通した形状を捉える Histograms of Oriented Gradients(HOG) [6] が 2005 年に提案された．HOG 特徴量は，画像を分割した各局所領域で作成した勾配ヒストグラムを 1 つのベクトルとしてまとめた特徴量である．物体の形状を表現する HOG 特徴量は，人検出以外にも有効であり，画像分類等の研究にも利用されている．しかし，高次元の特徴量であるため，直接的に画像全体から特徴を抽出する画像分類問題への適用が難しい．そこで，局所特徴量をベクトル量子化し，画像全体の特徴をヒストグラムで表現する Bag-of-Features(BoF) [5] というアプローチが提案された．その後，画像分類における特徴表現には，特徴ベクトルがもつ情報をより豊かに表現できるフィッシャーベクトル [16] や，特徴量を表現するためのメモリ使用量の削減を目的とした Vector of Locally Aggregated Descriptors (VLAD) [12] が提案された．

2010 年以降になると，handcrafted feature と呼ばれる SIFT や HOG のように研究者の知見に基づいて設計した特徴量に対して，学習により特徴抽出過程を自動獲得するアプローチとして Deep learning [10] が脚光を浴びている．handcrafted feature では，研究者の知見に基づいて設計したアルゴリズムにより特徴量を抽出・表現していたため最適な方法であるとは限らない．Deep learning では，認識に有効な特徴量の抽出処理を自動化する全く新しいアプローチである．Deep learning による画像認識結果は，一般物体認識のコンテストで圧倒的な成績を収めており，今後はさらに多くの分野での利用が期待されている．

このように特徴抽出の技術は，問題設定に合わせて最適な特徴量を設計し，より良い特徴を統計的学習手法で選択するというアプローチから，Deep Learning のような自動獲得するアプローチへと遷移しつつある．しかし，Deep Learning で自動獲得された特徴抽出過程は，計算コストが考慮されていないため，今後は研究者・開発者による計算処理の最適化が重要な課題となる．

4. 識別器（統計的学習手法）

1990 年代後半に入ると汎用コンピュータの進化に伴い，大量のデータを高速に処理できるようになったことから，画像から高次元の特徴量ベクトルを抽出し，統計的学習手法を用いて識別する手法が主流となった．統計的学習手法は，クラスラベルが付した大量の学習サンプルを必要とするが，ルールベースの手法のように研究者がいくつかのルールを設計する必要がないため，汎用性の高い識別器を学習できる．顔検出では Haar-like 特徴量と AdaBoost，人検出では HOG 特徴量と SVM の組み合わせが広く使用されている．2008 年に提案された Deformable Parts Model(DPM) [9] は，人全身を捉えるルートフィルタとパーツを捉えるパーツフィルタの位置関係を Star Model により表現する．そして，ルートフィルタとパーツフィルタの見えに対するスコアと，パーツフィルタの位置に対するスコアの和が学習サンプルにおいて最大となるように Latent SVM により学習する．これにより，DPM では姿勢の大きな変化に対応することが可能となった．

物体検出は対象クラスと非対象クラスに識別する 2 クラスの問題設定であったが，画像分類やセマンティックセグメンテーションのような応用では多クラスの問題設定が扱われるようになった．マルチクラス識別器の一つである Random Forest [3] は，ランダム性を取り入れたアンサンブル学習手法であり，2006 年以降に画像認識の分野で利用され始めている．2011 年に提案された Exemplar SVM [14] は，識別問題を正例 1 つに対して負例多数という部分問題に分割し，部分問題毎に SVM 識別器を構成するアプローチである．このような事例ベースのアプローチは計算コストが高いが，事例が持つ情報からセグメンテーションや 3 D モデル等を獲得することが可能であり，さらなる進展が今後期待されている．

このように識別器の技術では，既存の識別器をただ利用するだけでなく，その学習過程に工夫を取り入れて高精度化するアプローチへと遷移してきた．今後は，問題設定の多クラス化に伴い増加する学習時間，識別時間，学習サンプルの収集コスト等の学習における効率化を解決するアプローチが期待されている．

5. まとめ

本稿では，物体認識の動向について述べた．実用化につながる画像認識技術が大きな進展を遂げた要因の一つに，統計的学習手法と画像局所特徴量の進化が挙げられる．統計的学習手法と画像局所特徴量の組み合わせは高性能な認識を実現することができたが，位置ずれや誤ラベルがない良質な学習サンプルを大量に必要とする．このような大量な良質の学習サンプルの収集コストは非常に高いため，インターネットを利用したビッグデータの活用が望まれている．大量のデータを簡易に利用できる環境が整いつつある現在では，ビッグデータを対象とした研究に遷移してきており，ビッグデータ時代における新たな物体認識の問題設定が今後の鍵になると思われる．

Deep Learning は，3 章で述べたように大量の学習サンプルから特徴抽出過程を学習により自動獲得するアプローチであり，ビッグデータとの相性が良く，今後さらなる進展が期待できる研究分野である．一方で，Fine-grained visual categorization（詳細画像識別）という新しい問題設

定では，鳥の種類等の詳細なクラス識別を対象としており，クラウドソーシングを利用して人の知見を取り入れた認識アプローチが提案されている [8]．このような人の知見とのハイブリッドによる認識アプローチと Deep Learning による特徴抽出過程の自動獲得は対照的ではあるが，共に大変興味深く今後の進展に大いに期待したい．

顔検出や歩行者検出等の統計的学習手法を用いた物体検出は 2 クラス識別を対象としてきたが，ビッグデータの利用に伴い，対象とするクラス（カテゴリ）数がスケールアップしている．国際会議 CVPR2013 で Best paper award winner である Google の論文 [7] においては，100,000 クラスの物体検出が報告され，100K クラスの DPM を 20 秒以下で実行することに成功している．検出精度自体にはまだまだ課題が残されている印象ではあるが，先駆的な意義深い研究であり，今後の物体認識の発展におけるマイルストーンとして果たす役割は大きいものと考える．このような超多クラス問題における検出精度の向上は依然課題であり，全てのクラスに関して網羅的に良質な学習サンプルを集めるのはビッグデータの利用だけでは不可能であることから，今後は attribute や転移学習の概念が役に立つと考えられる．さらに，超多クラス識別では識別演算に要する計算コストを如何に低くするかが課題であり，ハードウェアに頼ることなく，アルゴリズムによる高速化，例えばイナリコード型局所特徴量やハッシングの利用が鍵となるであろう．

2000 年以降に実用化の芽がでてきた物体認識技術は目覚ましい進歩を遂げ，2010 年以降では大幅にスケールアップしている．今後，積極的なビッグデータの取り込みと認識アルゴリズムのさらなる進化により，新たな問題に対しても適応可能なアプローチに物体認識技術が進展していくことと期待している．

参考文献

[1] Ambai, M. and Yoshida, Y.: CARD: Compact and real-time descriptors, *International Conference on Computer Vision*, pp. 97–104 (2011).

[2] Bay, H., Tuytelaars, T. and Van Gool, L.: Surf: Speeded up robust features, *European Conference on Computer Vision*, pp. 404–417 (2006).

[3] Breiman, L.: Random forests, *Machine Learning*, Vol. 45, No. 1, pp. 5–32 (2001).

[4] Calonder, M., Lepetit, V., Strecha, C. and Fua, P.: Brief: Binary robust independent elementary features, *European Conference on Computer Vision*, pp. 778–792 (2010).

[5] Csurka, G., Dance, C., Fan, L., Willamowski, J. and Bray, C.: Visual categorization with bags of keypoints, *Workshop on statistical learning in computer vision, ECCV*, Vol. 1, No. 1-22, pp. 1–2 (2004).

[6] Dalal, N. and Triggs, B.: Histograms of oriented gradients for human detection, *Conference on Computer Vision and Pattern Recognition*, Vol. 1, pp. 886–893 (2005).

[7] Dean, T., Ruzon, M., Segal, M., Shlens, J., Vijayanarasimhan, S. and Yagnik, J.: Fast, accurate detection of 100,000 object classes on a single machine, *Conference on Computer Vision and Pattern Recognition*, pp. 1814–1821 (2013).

[8] Deng, J., Krause, J. and Fei-Fei, L.: Fine-grained crowdsourcing for fine-grained recognition, *Conference on Computer Vision and Pattern Recognition*, pp. 580–587 (2013).

[9] Felzenszwalb, P., McAllester, D. and Ramanan, D.: A discriminatively trained, multiscale, deformable part model, *Conference on Computer Vision and Pattern Recognition*, pp. 1–8 (2008).

[10] Hinton, G. E., Osindero, S. and Teh, Y.-W.: A fast learning algorithm for deep belief nets, *Neural Computation*, Vol. 18, No. 7, pp. 1527–1554 (2006).

[11] Huang, C., Ai, H., Yamashita, T., Lao, S. and Kawade, M.: Incremental learning of boosted face detector, *International Conference on Computer Vision*, pp. 1–8 (2007).

[12] Jégou, H., Perronnin, F., Douze, M., Sanchez, J., Perez, P. and Schmid, C.: Aggregating local image descriptors into compact codes, *Pattern Analysis and Machine Intelligence*, Vol. 34, No. 9, pp. 1704–1716 (2012).

[13] Lowe, D. G.: Distinctive image features from scale-invariant keypoints, *International Journal of Computer Vision*, Vol. 60, No. 2, pp. 91–110 (2004).

[14] Malisiewicz, T., Gupta, A. and Efros, A. A.: Ensemble of exemplar-svms for object detection and beyond, *International Conference on Computer Vision*, pp. 89–96 (2011).

[15] Mita, T., Kaneko, T., Stenger, B. and Hori, O.: Discriminative feature co-occurrence selection for object detection, *Pattern Analysis and Machine Intelligence*, Vol. 30, No. 7, pp. 1257–1269 (2008).

[16] Perronnin, F. and Dance, C.: Fisher kernels on visual vocabularies for image categorization, *Conference on Computer Vision and Pattern Recognition*, pp. 1–8 (2007).

[17] Rublee, E., Rabaud, V., Konolige, K. and Bradski, G.: ORB: an efficient alternative to SIFT or SURF, *International Conference on Computer Vision*, pp. 2564–2571 (2011).

[18] Viola, P. and Jones, M.: Robust real-time object detection, *International Journal of Computer Vision*, Vol. 4 (2001).

[19] 佐部浩太郎，日台健一：ピクセル差分特徴を用いた実時間任意姿勢顔検出器の学習，画像センシングシンポジウム (2004).

ⓒ 2016 Information Processing Society of Japan