基専

1 画像・映像の認識と理解のこれまでとこれから

佐藤真一（国立情報学研究所）

画像・映像の認識と理解：なぜ難しいのか

画像・映像の認識と理解は，視覚認知の機構の解明という知的興味から，実際の画像・映像の意味解析への要望という実用的要請まで，広範な目的のため検討されてきている．特に，監視カメラ映像の自動監視，ロボットのナビゲーション・インタラクションのための実際に機能し得る視覚機構の実現，自分の代理（エージェント）により大量の画像・映像から所望の対象を検出する究極のマルチメディア検索の実現等，昨今は特に実用面からの要請が強い．その一方で，画像・映像の認識と理解は困難な技術課題であると知られている．まずはその理由について考察しよう．

画像・映像は，テキスト・数値データと比較して，計算機でまともに扱えるようになったのはかなり最近のことである．人工知能テストである Turing test，テキストに基づく人工知能プログラム ELIZA や SHRDLU 等が発表されたのが 1950 ～ 60 年代であり，計算機処理のため著作権切れの書物を電子化しようというプロジェクト・グーテンベルクが立ち上がったのが 1971 年であって，計算機の黎明期にすでにテキスト処理はどんどん広がりを見せていた．その一方，同じ時期に計算機で画像を扱うのはきわめて大変であった．金出武雄カーネギーメロン大教授の 1973 年の京都大学博士論文[1]の研究は，画像入力から顔認識まで一貫して実現して見せた世界で初めての研究として認識されているが，まずは画像の入力のためにフライングスポットスキャナという機械を計算機に接続するための回路を自前で設計・作成し，かつ結果の画像を出力するため，蓄積型 CRT への画像出力用回路も自前で設計・作成する必要があった．これでは計算機による画像解析の研究の広がりは望むべくもなかったであろう．さらに，画像・映像はデータとして巨大であり，計算機で扱うのはそれだけ困難である．ブログ等で 1MB に達する文書（日本語テキストであれば 400 字詰原稿用紙 1,310 枚以上）を書くのはかなり骨であると考えられるが，デジカメで写真を撮っていると数十 GB のメモリがすぐにいっぱいになるし，ハードディスクレコーダでは 2 ～ 3TB の容量がすぐに埋まってしまう．歴史的に考えても，計算機用の安価な外部記憶装置として CD-ROM が出現し，画像と低品質の映像（Video CD）が使えるようになったのが 1980 年代，DVD が出てきて映像がまともに使えるようになったのが 1990 年代であり，GB ～ TB オーダのフラッシュメモリまで出てきて画像・映像が自由自在に扱えるようになったのはごく最近のことである．

こうした技術的な困難さとは別に，画像・映像の認識と理解には本質的なむずかしさがある．テキストや数値がそもそも人工的なデータであり，たとえば単語などはそのままその単語の持つ意味と関連するのに対し，画像・映像は実世界をそのまま観測した「生」の情報であり，画像の各画素の色は正確に RGB 値として表現できているが，物体に対応する領域はどの画素が対応するのか（単語に相当）などは「見えない」．こうした問題は，観測可能な画素値などの情報と必要な意味レベルの情報とのかい離から，セマンティックギャップと呼ばれる．加えて，画像・映像の認識と理解は，人間にはあまりに簡単であるため，困難な問題ではないと当初（かついまだに）誤解されたという経緯もある．画像を見

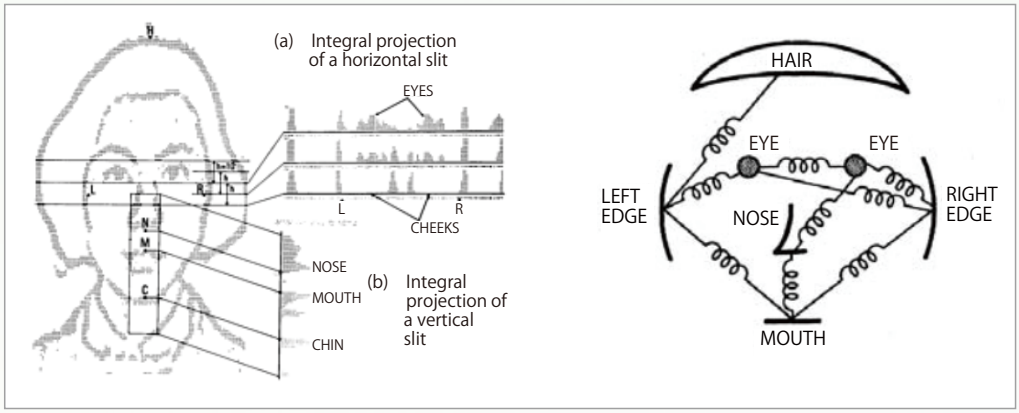

図 -1　顔画像認識の例：目と鼻と口の位置関係がこうだから…とプログラミングしようとした（左：文献 1）右：文献 2））

て，そこにイヌが写っていると判断するのに苦労する人はいないが，計算機にはとても難しい．人工知能研究の巨匠マサチューセッツ工科大の Marvin Minsky 教授は，1966 年のある日，大学院生を呼び出し，夏休みの宿題に，コンピュータにカメラをつなぎ，シーンを説明するプログラムを作成せよと命じたという．画像認識の問題が学生の夏休みの問題にちょうどいいと考えたのだ．一方，当時はチェスを指すプログラムを実現することこそが人工知能実現の王道であると考えられた．ところが，人間のチェスチャンピオンは計算機に敗れてしまったが，いまだにイヌの画像を間違いなく判断できる計算機プログラムは実現できていない．また，これは認知心理学の課題だが，人間にも人間自身がどうやって画像の認識を行っているのか分からないという点も問題である．

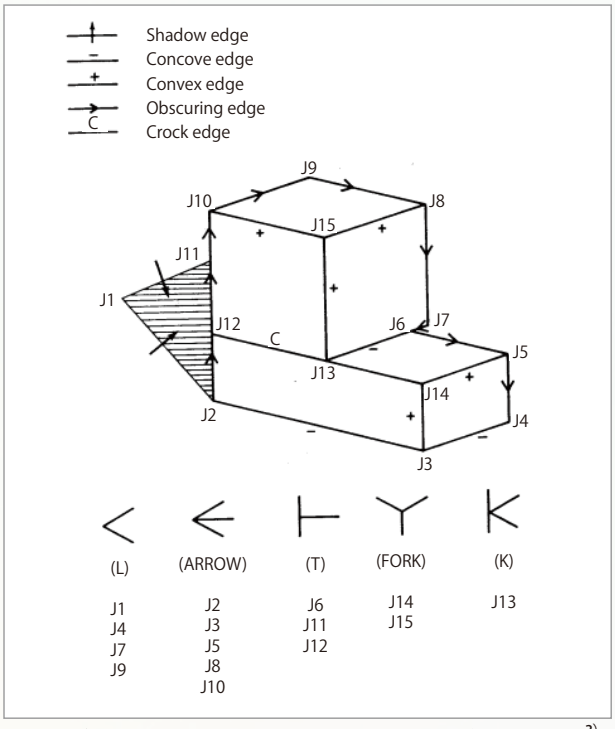

図 -2　ブロックシーンの認識：ルールによりシーンが認識できた [3]

画像認識研究の黎明

こうした中，画像認識研究はどのように立ち上がっていったのか．1970 年代の黎明期には，まずは人が画像を認識するようにプログラミングするというアプローチがとられた．顔の認識では，先の金出教授の博士論文 [1] や文献 2）等の先駆的な研究が挙げられる（図 -1）．文献 1）では顔部品の検出を二値化画像の射影などアドホックな方法で実現しているが，顔全体での制約と顔部品検出との相互作用や，文脈などの利用が試みられており，文献 2）でも顔部品間の位置関係の制約をばねモデルで記述しており，興味深い方法が提案されている．しかし，そもそも人の認識過程の説明が困難である上，かつさまざまな顔に対応するためにいちいちプログラムを変更する必要があり，限界を迎える．一方，Waltz は積み木のようなブロックのシーンを表す線画の認識のため，線分が満足するべき制約をすべてルールとして計算機に搭載し，制約充足問題としてシーン認識を実現することに成功した [3]（図 -2）．この成功を受けて知識をルールとしてシステムに搭載し，画像認識を人工知能の問題として解く方法が広まり，自然画の認識まで実現された．Brooks の発

表した万能 3D シーン認識システム ACRONYM[☆1] は，ルールさえ搭載すればどんなシーンでも認識できるとされたが，そもそも必要なルールを記述する困難さが判明した．これは AI におけるフレーム問題そのものであり，こうしたアプローチはとん挫してしまう．一方，やはり人間の認識過程に基づいて手法を設計しようという試みもあり，認知心理学の知見に従い，ゲシュタルト[☆2] やアフォーダンス[☆3] を考慮に入れた画像認識手法や，写っている物体の機能に着目して認識しようという機能モデルも検討されたが，結実しなかった．この後，画像解析研究者は画像認識研究から離れ，ステレオ計測等画像に基づく計測に注力することとなり，画像認識研究は急速に衰退していく．

機械学習としての画像・映像意味解析

1990 年代に入り，顔ならびに文字認識において新たなアプローチが奏功しだす．大量の顔や文字の画像を集め，ニューラルネットワークなどの機械学習により認識問題を解くアプローチがとられ，成功を収めた．このアプローチのポイントは，どのように画像認識が機能しているかはまったく問わない点である．このようにして世界初の実用的な精度の顔検出器を実現した例が Rowley らによる二並列のニューラルネットを用いた顔検出手法であり，大量の顔画像を集めた CMU-MIT データセットを構築し利用している[☆4]．この考え方は一般の物体へと拡張され，一般物体認識のためのデータセットが構築され，研究に供されていった．その例として，COIL, Caltech 101/256, PASCAL VOC 等が挙げられ，Bag of Visual Words 等の画像表現やそれに基づく機械学習アルゴリズムの研究が一気に進んだ．映像意味解析・検索においては，米国標準技術局主催の TRECVID による数百～数千時間規模の映像データが整備され，画像に続き映像の認識と理解の研究も顕著に進んできている．

その一方で，認識対象の物体種別（カテゴリ）の選択が問題となってきた．上記の Caltech 等では，研究者らがあらかじめ選んだカテゴリが用いられたが，イヌやネコというカテゴリはないのにムカデやサンヨウチュウというカテゴリがあるなど，その恣意性が問題となってきた．ImageNet[☆5] では，カテゴリを概念辞書 WordNet から網羅的に選ぶことによってこの恣意性の問題を排除し，数万という大量のカテゴリに基づく画像意味解析用データセットが実現されている．

新たな潮流

意味の問題の深みへ

機械学習に基づくアプローチでは，各カテゴリはあくまでラベルとして客観的に扱い，イヌやネコであってもラベル -A やラベル -B として扱っていた．しかし，特に ImageNet のように数万カテゴリを扱おうとすると，なかなか高精度の認識が難しくなってきて，その裏の意味が無視できなくなってきた．たとえば ImageNet で「アカアシシギ」と「カラフトアオアシシギ」は独立したカテゴリだが，その厳密な識別は大変困難である．そこで，カテゴリ間の概念的関連性を明示的に扱おうというアプローチが出てきた．Smith らは，TRECVID データを対象とし，関連する概念の学習データをそれなりに利用して学習データの不足を補う手法を提案した[☆6]．カテゴリそのものを識別対象にするのではなく，カテゴリ間で共通する属性（Attribute）を識別対象にし，その識別結果で元のカテゴリの認識精度の向上を図る方法も提案された[☆7]．文献 4) は，関連するカテゴ

☆1 Brooks, R. A. : Symbolic Reasoning Among 3-d Models and 2-d Images, Articial Intelligence, **17**, pp.285-348 (1981).
☆2 対象を個別に捉えるのではなく，全体として捉えようという心理学の考え方．
☆3 人が対象をどのように使うかという関係性のこと．ただしこれは大変広く使われている誤用であるといわれている．
☆4 Rowley, H. A., Baluja, S. and Kanade, T. : Neural Network-based Face Detection, Proc. of Computer Vision and Pattern Recognition, pp.203-208 (1996).
☆5 http://www.image-net.org/
☆6 Smith, J., Naphade, M. and Natsev, A. : Multimedia Semantic Indexing Using Model Vectors, Multimedia and Expo, IEEE International Conference on, **2**, pp.445-448 (2003).
☆7 Ferrari, V. and Zisserman, A. : Learning Visual Attributes, NIPS (Eds. by Platt, J. C., Koller, D., Singer, Y. and Roweis, S. T.), Curran Associates, Inc. (2007).

リをたどり，未知のカテゴリ（学習データのないカテゴリ）の識別を行ってみせた（図 -3）．認識精度とは別のターゲットとして，大量にカテゴリがあると識別処理が遅いという問題も着目された．広く使われている識別技術は 2 クラスの識別問題を扱うものであり，これを複数カテゴリの識別問題に適応する場合には，one vs all や one vs one 等の方策により 2 クラス問題として解決するのが主であるが，カテゴリ数が N の場合，one vs all の計算量は N, one vs one の場合は N^2 の計算量となってしまう．これに対し，Label Embedding Tree ならびに引き続く多くの研究では，カテゴリ間の関連性から全カテゴリを木構造に再構成し，計算量を $\log N$ に抑える方法を提案している☆8．

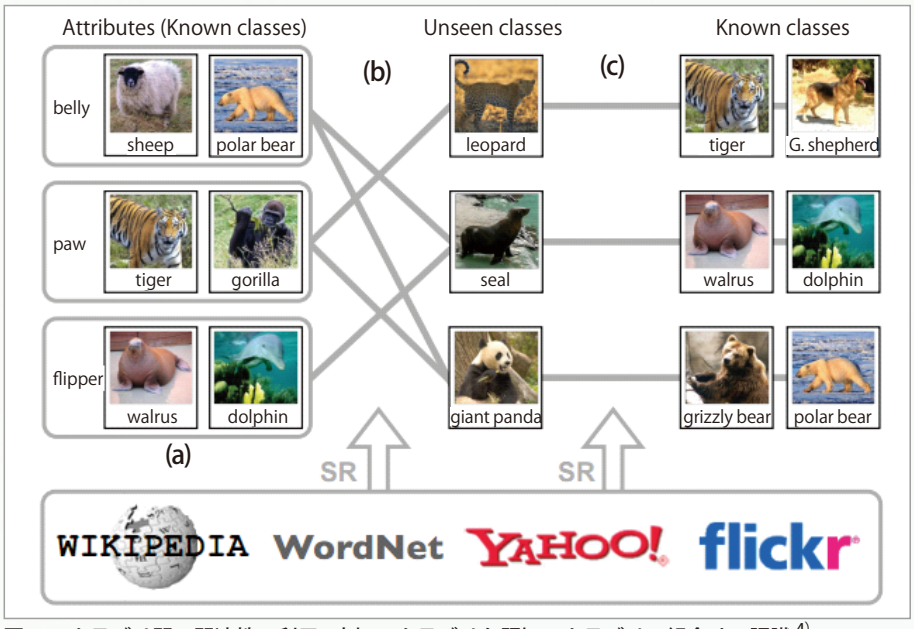

図 -3　カテゴリ間の関連性の利用：未知のカテゴリも既知のカテゴリの組合せで認識 [4]

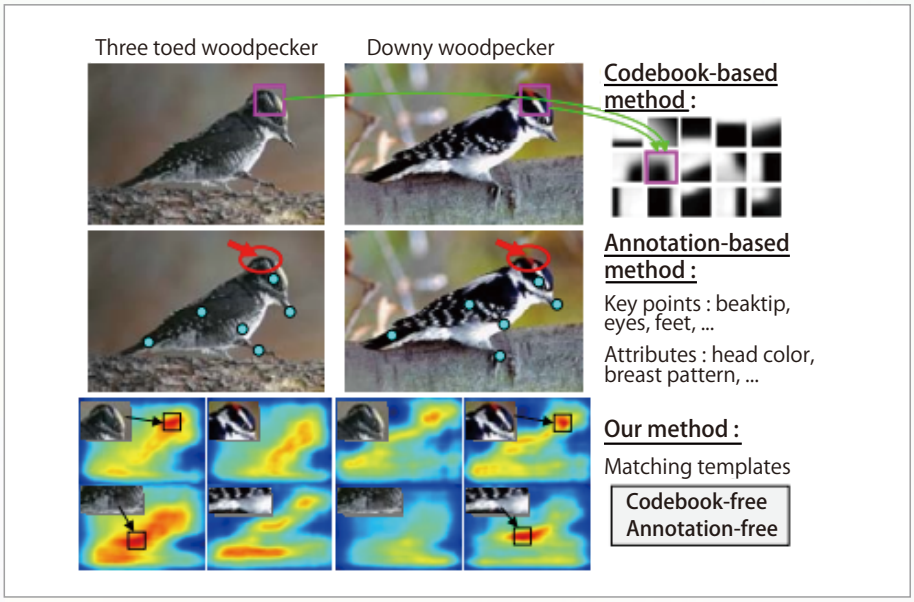

図 -4　鳥の詳細な識別：ミユビゲラ（Three toed woodpecker）とセジロコゲラ（Downy woodpecker）が識別できるという [5]

概念数が増えてくると前述のように識別困難な概念集合が出てくる．加えて，識別能力の高かった機械学習技術が相対的に弱体化し，k 近傍法を用いた認識手法等と性能が変わらなくなってくるという．その一方，画像中の対象が鳥だと分かって識別すると詳細な種の識別までが可能であり，Fine-Grained Visual Categorization として検討されている [5]（図 -4）．一般の物体の識別とは異なり，くちばしの色とか斑点の有無など識別過程が説明できることも一因と考えられる．

1980 年代に，人工知能という意味の問題と決別し，1990 年代にも機械学習の導入により意味の問題に一定の距離を置いていたところ，ここにきて意味の問題に立ち返らざるを得なくなっているようにも見え，興味深い．

☆8 Bengio, S., Weston, J. and Grangier, D. : Label Embedding Trees for Large Multi-class Tasks, Proc. of NIPS (2010).

図 -5　アフォーダンスの利用：シーンに対し可能な人間のインタラクションを推定した例 [6)]

ニューラルネットワークの逆襲

ニューラルネットワークに基づく深層学習（ディープラーニング）が注目を集めている．詳細については本特集の記事「ディープラーニングによる画像認識」を参照いただきたい．ディープラーニングにより，画像意味解析，顔認識，情景文字認識等で記録がどんどん塗り替えられており，人間の認識性能も凌駕しつつあるという．最近の画像・映像認識研究でも，識別器を畳込みニューラルネットワーク（DCNN）に変えただけで顕著に精度向上するという報告が多く見られる．上記の記事では，その限界についても論じられているが，まだまだ「のびしろ」のある興味深い技術であることは間違いない．

人による認識と計算機による認識

1980 年代における認知心理の知見を用いようというアプローチは事実上結実しなかったが，これに類する試みも昨今見られる．たとえば人間による画像意味解析の特性と計算機による特性との共通点や違いを解析しようという試みや，またこれにより計算機による画像意味解析で注力すべき問題点の洗い出しを図る試みがなされている☆9, 10．本特集の記事「画像識別と画像復元」では，画像識別の逆問題としての画像復元について解説しており，特に画像識別の内部で何が起こっているのかを画像化することにより人間に理解させようとしている．画像の意味理解において計算機内部の処理についてはブラックボックスとして機械学習に任せてしまおうというアプローチとは逆であり，こうした試みも興味深い．TRECVID の Multimedia Event Recounting（MER）というタスクでは，計算機が出力した映像意味解析結果に至る過程を計算機により説明させることが目的であり，上記の画像復元とも通じる．ただし，TRECVID MER タスクでは，主としてテキストで説明させようとしている．

1980 年代に検討されたアフォーダンスの利用についても再び検討されている．文献 6）では，物体と人間とのインタラクションを解析し，それに基づいて物体の種別やシーンの解釈を行おうという試みであり，まさしくアフォーダンスを利用した画像・映像の認識と理解である（図 -5）☆11．これが可能になったのは，1980 年代に比べて物体の検出や人体の検出・追跡技術の性能が圧倒的に向上したことが考えられ，アフォーダンスなどの考え方が実際に実装可能になってきたためと考えられる．1980 年代に検討されたほかの方法についても，再考の余地があるかもしれない．

画像・映像の新たな使われ方：関連情報の利用

従来，画像・映像の認識と理解では，与えられた画像や映像に対し，あたかも人間が行うような解釈を行うのが聖杯（holy grail）であり，研究の王道で

☆9 Borji, A. and Itt, L. : Human vs. Computer in Scene and Object Recognition, Proc. of CVPR (2014).
☆10 Parikh, D. : Recognizing Jumbled Images : The Role of Local and Global Information in Image Classication, Proc. of ICCV (2011).
☆11 文献 6）のほかにも Grabner, H., Gail, J. and Gool, L. V. : What Makes a Chair a Chair?, Proc. of CVPR (2011) など．

あると考えられた．一方で，一般ユーザの画像・映像へのかかわり方が明らかに変質してきている．スマホなどで気軽に撮影し，インターネットにアップロードし，友人らとシェアするなど，コミュニケーションの一部に組み込まれている．特にSNS（ソーシャルネットワークサービス）の利用がその最たるものであり，その場合には画像・映像には撮影日時，撮影場所，撮影者等の重要な付加情報が付随することになる．本特集の記事「ソーシャルネットワーク上の画像を認識・理解する」では，こうした状況について詳細に述べられているので参照いただきたい．こうした情報に基づき，ある人物が撮影した一連の画像を解析するだけで，その人物は実はアジア系の女性であり，ニューヨークで女性の友人同士でショッピング中，等が分かるという．

また，我々は日々検索エンジンを利用している．検索エンジン運営側からすれば，我々のクリックする情報を大量に集めれば，どの問合せに対してどのような検索結果を提示した場合にはどれがクリックされたかという情報が大量に集まることになり，とりもなおさず問合せに対する学習データとして利用できることになる．こうした情報はclick throughデータと呼ばれ，利用者により無料で無数提供される付加情報であり，これに基づく画像・映像の認識と理解の研究も進められている．

今後の展望

学術研究としては，意味の問題への取り組みが重要と考えている．これは，画像・映像の意味解析に比べて突出して研究の進んでいる自然言語理解でもやはり困難な問題であり，完全に解決するとは考えられない．しかし，不特定多数の人と対話をするようなシステムのための視覚の実現等においては，たとえば認識するカテゴリが事前にすべて決まっているわけではないような，本質的に「開かれた」システムが必要となり，あらかじめ学習データを整備することが困難になると考えられる．この場合には意味の問題へのある程度の決着が必要となろう．困難も予想されるが，認識過程の可視化としての画像復元は重要なツールとなり得る．一方，対象を閉じた問題として捉えることができ，かつ大量の学習データが用意できる場合には，意味の問題を避けた，従来通りの機械学習的なアプローチが有効と考えられる．インターネットの検索エンジンは，対象は巨大ではあるが閉じており，click throughデータも利用できる．また，SNSもユーザらの振舞いのほとんどが観測可能と考えると閉じた世界であり，こうしたアプローチが有効と考えられる．このような，大量の学習データを用いた教師あり学習が可能な状況では，特に高精度が達成可能なディープラーニングの利用は重要と考えられる．

参考文献

1) Kanade, T. : Picture Processing System by Computer Complex and Recognition of Human Faces, Ph.D. Thesis, Kyoto University (1973).
2) Fischler, M. A. and Elschlager, R. A. : The Representation and Matching of Pictorial Structures, IEEE Trans. on Computers, **C-22**, 1, pp.67-92 (1973).
3) Waltz, D. L.: Understanding Line Drawings of Scenes with Shadows, in Winston, P. H., ed. : The Psychology of Computer Vision, McGraw-Hill (1975).
4) Rohrbach, M., Stark, M., Szarvas, G., Gurevych, I. and Schiele, B. : What Helps Where—and Why? Semantic Relatedness for Knowledge Transfer, Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, pp.910-917 (2010).
5) Fei-Fei, L., Yao, B. and Bradski, G. : A Codebook-free and Annotation-free Approach for Fine-grained Image Categorization, Proc. of CVPR (2012).
6) Delaitre, V., Fouhey, D. F., Laptev, I., Sivic, J., Gupta, A. and Efros, A. A. : Scene Semantics from Long-term Observation of People, Proc. of ECCV (2012).

(2015年4月28日受付)

佐藤真一（正会員） satoh@nii.ac.jp

1987年東京大学工学部電子工学科卒業．1992年同大学院工学系研究科情報工学専攻博士課程修了．学術情報センター助手等を経て，2004年より国立情報学研究所教授，現在に至る．1995～97年まで，米国カーネギーメロン大客員研究員としてInformedia映像ディジタルライブラリの研究に従事．工博．画像理解，画像データベース，映像データベース等の研究に従事．