Vol.2016-CVIM-200 No.16
2016/1/21

コンピュータビジョンはビジョンを超えられるか －CV の発展を概観し、今後への期待を述べる－

白井良明[†1]

概要：これは、CVIM 研究会の 200 回を記念した講演の概要である。まず、これまでの発展を概観し、現在の CV の傾向と問題点を述べる。とくに、最近盛んになっている学習による画像認識の限界を指摘する。人の能力に近付けるためには、まだやることがある。そこで、人の視覚能力実現のため、のアプローチとして、これまで筆者が試みてきた研究をいくつか簡単に紹介する。

キーワード：コンピュータビジョン、ブラックボックス化、学習、ＣＮＮ　非階層的手法、default model

Can Computer Vision Exceed Vision? －Review Development of CV and Describe Future Expectation－

YOSHIAKI SHIRAI[†1]

Abstract: This manuscript is summary of the talk at the 200th meeting of CVIM. First, the development of CV is reviewed, and then the tendency and problems of current CV are described. Especially learning methods for CV are picked up and the current limitations are pointed out. There are many things to be done for approaching human ability. As an approach to human vision ability, several research methods are briefly introduced which the author has attempted.

Keywords: computer vision, black box, learning, CNN, heterarchical method, default model

1. はじめに

今回の講演は、情報処理学会の CVIM 研究会 200 回記念として依頼された。なぜ私が行うかを疑問に思われる方のために、簡単に理由を推察してみる。

CVIM の源流は、1973 年に設立されたイメージプロセッシング研究委員会である．当時は 3 次元シーンを直接コンピュータに入力できる研究機関は世界でもわずかであり、大部分写真をゆっくり入力していたのがこの名前の由来であろう。この委員会は，研究会に設立の準備のための非公開の委員会であった．主査は、当時京大教授の坂井先生、幹事は京大助教授であった長尾先生と電総研にいた私であった。3 年後に公開のイメージプロセッシング研究会となり、当時の学会の規定により、4 年後に名称を変更してコンピュータビジョン研究会となった。この CV 研は 1996 年に CVIM と名称を変更するまで 16 年間続いた。この間、研究会委員、幹事、主査を務めた。研究会は 2 カ月に一度で、発表件数も少なかったので、発表時間は長く、質疑応答も厳しいといわれていた。

比較的長くコンピュータビジョンに携わった経験に基づき、これまでの発展をやや批判的に振り返るとともに、現状の問題点や今後の期待を簡単に述べる。なお、最新の状況に関しては知らないことも多いので、的外れと思われる部分もあろうが、話題提供とみなしてご容赦願いたい。また、おかしいと思われる点をご指摘いただければ幸いである。なお、比較的知られている文献は、参考文献から省いている。

†1 立命館大学グローバル・イノベーション機構
The Research Organization of Global Innovation、Ritsumeikan University

2. CV 研から CVIM 研へ

ＣＶ研の発足当時は、画像をコンピュータで処理すること自体が簡単でなかった。そこで、研究会の委員会では、最初に入力装置の紹介や説明を 2 年くらい行った。その後、処理を取り上げ、その説明とともに、研究会の幹事である田村氏が音頭をとって、多数で画像処理プログラムを開発した。これは、SPIDER という名前をつけ、大学に頒布するとともに、会社へは販売を行った。CV 研究が普及してくると、このような委員会活動はなくなり、研究会の運営に力を入れるようになった。

CV 研究の初期は、入力媒体はモノクロの画像であった。屋外風景認識にはカラー画像が有効であったが、入力と蓄積の問題からカラー画像を利用できる研究者は限られていた。

距離入力のために、ステレオ視の研究が行われていたが、信頼性が十分でなく、ステレオ視で得た距離情報を利用するには至っていなかった。電総研では、独自に開発したレンジファインダーによって得た距離情報から物体認識を行っていた。距離情報を使えたのは 1 年遅れてレンジファインダーを開発したスタンフォード大だけであったので、距

離情報からの物体認識は類似研究がなく、その成果は世界に知れ渡った。

コンピュータの進歩により動画も蓄えられるようになると、動画像処理の研究が盛んになった。以上のように、CV 初期は、入力情報の拡大がその発展の原動力となっていたともいえよう。

計算能力の向上による処理方法の拡大も発展の大きな要因といえよう。初期には、Hough 変換や領域分割も計算量が多すぎるとみなされた。それが、Active Contour や Optical Flow の計算のように、正則化を含む繰り返し計算も使えるようになった。

CV 研究者は増え、基本的な処理手法も増えてきた。CVIM は、ある程度飽和した CV 研究に一区切りつけ、応用を広げることにより、新たな発展を期待したともいえよう。

最近は、種々の手法をインターネットから入手できるので、プログラミングが簡便になったようである。さらに、SVM, 深層学習、AdaBoost のような学習も盛んに利用されるようになった。簡単な画像認識処理は、適当な入力と学習画像を与えれば、達成できるようになった。　これで、計算量さえ問題にしなければ、学習で何でも解決できると考える人も増えている。

以上、これまでの発展を主観的に概観したが、以下にもう少し具体的に述べる。

3. 手法の発展とブラックボックス化

コンピュータビジョンは、処理対象、入力の種類、アルゴリズムなどによって多様であるが、処理の複雑さと計算に関していくつかの処理を位置づけると図 1 のように近似できる。

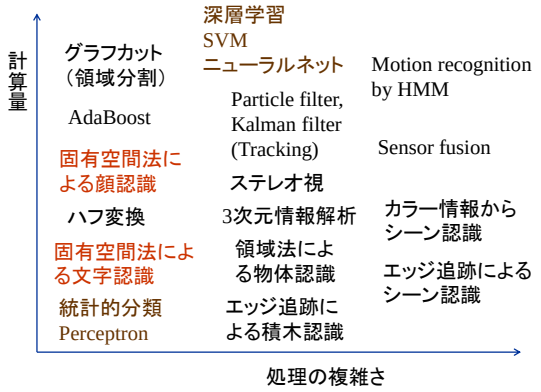


図 1　コンピュータビジョンの処理の性質
Figure 1　The characteristics of vision process

処理の複雑さは、プログラムを作るのにどれだけ労力がかかるかを表している。たとえば、1960 年代に文字認識に固有空間を用いた方法(PCA)が試みられた。その手法自体はさらに古くから提案されていて[1]、プログラミングは複雑ではない。ところが、当時のコンピュータのメモリと計算速度の制限から、文字のような小さい画像に制限された。この計算量の制限が緩められた時、Pentland が顔認識に適用した。これも処理自体は複雑ではない。

なお、Pentland は、CG でも優れた成果を出している。物体のモデルを有限要素法で表し、リアルな形や動きを表現した[2]ことで有名になった。有限要素法は、機械工学などで物体にかかる力を計算するために使われている。このように、工学の基本技術を情報処理に利用していたが、新しいと思われる手法も他の分野で開発されたものが多い。

PCA による顔認識は、顔が切り出された（セグメンテーション）場合には有効であるが、シーン内のどこにあるかが決まっていないと非常に効率が悪いため、実用には向いていなかった。それでも、PCA は簡単に使えるため、一時は流行した。しかし、一般的なシーン内の対象を認識するためには、CV で最も困難であるといわれるセグメンテーションのプログラムを書かなければならないため、適用範囲が限定されている。成功したのは、部分特徴の組み合わせとして顔を検出し、検出と同時に顔の部分も認識する方式である（例えば、有名な Viola-Jones の手法）。

動画像から人を追跡することは、原理的には複雑な処理が必要である。すなわち、人の形の変化、回転、隠れなどに対処することは簡単でない。一方、簡単な特徴を手掛かりとして、Particle Filter を用いた追跡を行う例が多い。Particle Fitler 自身は、追跡対象を見失ってもロバストに追跡するすぐれた手法である。しかし、単にそのプログラムが簡単に手に入るという理由で、隠れなどがない単純な追跡に使っていることが多い。Particle Filter 自体は与えられた問題の一部に適用できる手法であるので、CV の立場からは、問題をどのように定式化するか（どの特徴を用いるかなど）を重視すべきであろう。

最近、画像のセグメンテーションにグラフカットを用いることが多い。セグメンテーションの問題を、正則化を含む最適化問題として定式化し、OR で開発された効率のよい計算アルゴリズムを使い、高速化を図るという優れた手法である。CV の観点からは、グラフの辺のコストをどのように決めるか、また正則化の係数をどうするかが重要で、それによって結果が大きく異なる。

筆者らは、手ぶれするカメラで撮影したゴルフスイング動画から、人のシルエットを求める研究を行ったことがある。カメラの動きを補償する画像に変換し、各画素の時間変化に対する画素値にヒストグラムに基づく方法を提案した。論文原稿を提出したところ、同じことはグラフカットによっても可能であるという査読結果が返ってきた。グラフカットは最適化の手法に過ぎず、本質は別であることが理解されていなかったようである。しかたがないので、グラフカットを適用し、簡単な特徴を使うのでは不十分であ

り、初期の提案に基づく方法と組み合わせることが必要なことを示した[3]。

ところで、前述の処理の複雑さは、使えるツールに依存する。最近は多くの手法が手に入るようになり、プログラミングの負担を軽減している。課題が簡単ならば、それで十分であるが、少し複雑になると、十分な結果が得られないことが多い。このことは、PCA、Particle Filter、グラフカットにもいえる。たとえば、グラフカットの正則化の係数によって結果が大きく変わるが、その決め方に対しては、確立した方法はわからない。

一般に、まとまりのある処理をツールとしてブラックボックス化することは技術の進展に有効であると考えられている。とくに、電気回路や機械の機構などではブラックボックス化が有効に使われている。しかし、情報処理では入出力関係が複雑なため、すべてに有効で、有用なプログラムはほとんどない。例えば、音声認識用の HMM は入手できるが、画像処理に適した HMM は自分で作らなければならない。

4. 学習は CV の課題を解決できるか

学習は人の知能に迫るアプローチとして期待が大きい。学習は広い意味をもち、パラメータ調整、特徴の選択などには欠かせない。ここでは全体の一部をがくしゅうするのでなく、ニューラルネットのように、まとまりのある処理を学習することを取り上げる。それらは、入力（画像）全体を対象とすることを原則としている。

たとえば、前述の PCA による文字認識は、画像に 1 つの文字だけがあることを想定している。したがって、画像内のどこにあるかわからない顔の認識では、sliding window（顔の大きさのウィンドウを画像内で走査）が必要であった。これは、計算速度の向上によって可能となったが、性能は十分でない。

著者も学生の希望に応えるかたちで、種々の特徴を用い、学習によって数種類の乗り物、建物、動物、花を画像から抽出する研究を行ったことがある[4]。既存の特徴を使い、とくに変わった手法を開発してはいなかったが、多くの画像検索の研究者と同様に、いくつかの画像データベースに対して適用すると、対象物体の抽出率と精度はいずれも 80%以上と、以外にもよい成績となった。

しかし、根本的な問題が残されていた。例えば、図 2 は飛行機の抽出例である（画像の一部だけを表示）。中段右端は誤抽出の例、下段右端は抽出できなかった例である。前者には飛行機の一部と紛らわしい部分があり、後者はテクスチャのため、外形が得られていないのではないかと推察できる。このように誤りが与えられると原因が推察できる場合もある（もちろん、推察できない場合もある）。しかし、その対策は簡単ではない。適切な学習例を多く用いればいいかもしれないが、何をどれだけ学習したらよいかがわからない。

図 2　飛行機の抽出結果
Figure 2　Airplane extraction result

このような学習に用いる画像特徴としては、SIFT, CHLAC, Bag of Features, HOG などが提案されている。これらが有効であるのは、普通の画像は対象に依存する拘束があることが多いので、その一部から全体を推定できることである。単に、これらのヒストグラムに基づいて認識ができるのは、自然界の拘束によるものであり、運が悪ければ、失敗することも当然である。したがって、その性能を理論的に予測することは困難であろう。

最近、多層のニューラルネットワークが盛んに使われている。それに深層学習という魅力的な名前を付けている。画像処理に適した CNN(Convolution NN)は、簡単に入手できるので、とくに CNN の利用が多い。これは、複数の 2 次元の入力（画像）に、場所に依存しない convolution を行い、その結果を縮小している。この処理を数段繰り返した後で、全結合の層を複数設けている。このように、1 層あたりの学習する重みの数を減らすことにより、多層でも学習が高速にできる。入力は、対象そのものを使え、人が特徴抽出を行う必要のないことが利点となっている。さらに、120 万枚の訓練データを用いて 1000 クラスを分類するコンテストで首位となったことから、その性能がよいと信じられている。

しかし、CNN にも以下のような限界がある。

1） 大量の訓練データを必要とするが、普通はタグ付けされた訓練データを大量に得ることは困難である。また、どれだけのデータが必要かわからない。子供でも、絵本で見た少数の象の例から、動物園の象を認識するように、少ない訓練データから学習できる人の能力にはまだ達していない。

2） もし、対象が大きい画像の一部にある場合は、そのまま学習することができないか、時間がかかりすぎる

ので、あらかじめ別に位置決めをするか、スライディングウィンドウを適用するかの対策が必要がある。もし複数の種々の対象が含まれている画像を認識するためには、それぞれを独立に抽出しなければならず、互いの拘束を使った効率のよい認識になじまない。

3） 他の学習と同様に、誤りの対策が得られない。したがって、自動車の自動運転のような誤りが重大な損失をもたらす場合には、そのまま使うことができない。アルゴリズムの開発の段階で機械学習を用いることはあっても、車に搭載するには人が納得しなければならない（MIRU2015 における金道敏樹氏の特別講演の質疑応答より)。学習結果を人が参考にしてロバストな処理を作ることになる。それでも、人に参考になる処理を提案するという意味では有用であろう。

以上の限界に対して、ある程度対策が考えられているが、まだ根本的に解決するまでには至ってないようである。この限界は、複雑であろうと考えられる学習を、単純な最適化問題として解いていることにあるかもしれない。

5. 人の視覚への接近

CV が進歩したといえども人の視覚から学ぶべきことはある。その長所を取り入れ、実用化にはコンピュータの長所を生かした方法を適用することが望ましい。これに対する筆者の試みを簡単に紹介したい。

(1) 入力情報の拡大

電総研に入った 1969 年頃は､米国でロボットビジョンが研究されていた。そこでは、濃淡画像から積み木を認識することが主な課題であり、私もその流れで課題に挑戦することになった。ところが、照明条件によっては、異なる方向の面の明るさの差が少なく、ノイズとの区別がつかないことが多いことがわかった。つまり、原理的に不可能な問題を解こうとしていた。

そこで 2 つの方法を考えた、１つは、照明方向が異なる 4 種類の画像を用いる方法である[5]。これは、人は見にくいときは照明を変えてみることに相当する。もう１つは、光切断法によるレンジファインダーの発明（特許は切れている）とその利用[6]である。人は両眼立体視によって直接立体情報を得ているが、コンピュータによるステレオ視に信頼性が低かったので、機械に適した方法で入力情報を拡大し、人に近い入力情報を得ようとしたのである。

(2) 入力に関する情報不足への対処

濃淡画像から特徴を抽出する場合、普通は閾値を用いて特徴を決定する。ところが、入力の性質が十分わかっていない場合は､適当な閾値をきめることができない｡さらに、画像の場所によって閾値を変える必要がある場合には対処できない。

このような対処の最初の例は、MIT で行った積み木の認識である。照明条件や積み木の種類や配置に関する事前知識がない場合、従来は適当な閾値でエッジを抽出して線画を作り、それを積み木として解釈していた。しかし、図３の中央のように、望ましい線が抽出されないことが多い。そうすると、その結果に基づいて認識を行うという方法は破たんする。これは、Minsky が従来の Hierarchical method ではどこかで誤ると、それが最後まで伝播するので、信頼性が低くなると主張していた。

筆者は、まず、明らかなエッジを抽出し、線画の一部を作り、それを部分的に解釈し、その結果に基づいて仮説を立て、検証する形でより信頼性の低い手掛かりを探すという方法をとった[7]。この方式は、どの順番で線が見つかるか、またどの順で積み木が認識されるかは、あらかじめ決められておらず、プログラムはプロダクションシステムと似た制御となる。この方式では、同図右のように、望ましい結果が得られることが多い。この方式は、Minsky にもオンラインのデモをすることができ、満足してもらえた。同じ AI Lab の Winston を含め MIT の研究者から世界に知らせてたいただいた。

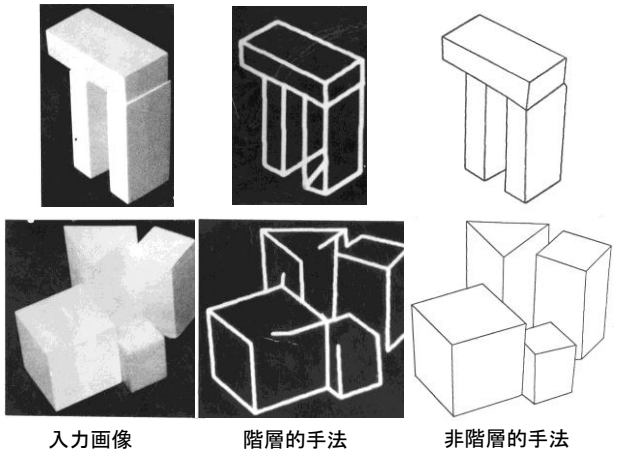
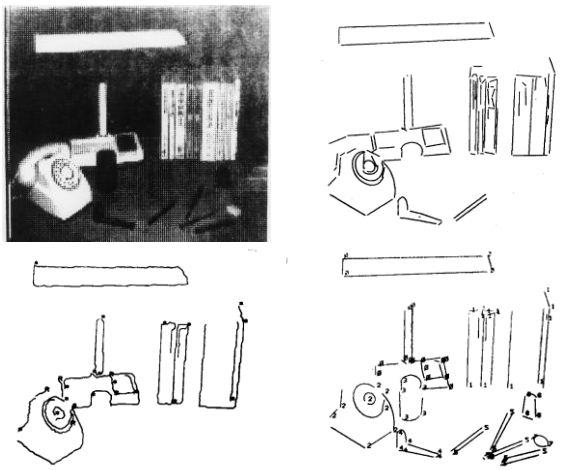


図 3　非階層的手法と従来手法の比較
Figure 3　Performance of heterarchical method

この方式は、電総研で同じ研究室の杉原によって距離データによる多面体認識に拡張された[8]。

(3) 一般物体への拡張

上記の方式は対象は何でもよいので、帰国後、大型プロジェクトの一部として物体認識を受け持った時、より一般的なシーンとして、普通のオフィスの机上シーンに拡大した。積み木は黒い背景におかれていたので、問題が簡単であると思われるのは心外であったので、図 4 のような灰色の机により明るい物体やより暗い物体が置かれているシーンを認識した[9]。



左上：入力。右上：単なる線画抽出結果。左下：最初に得られた線画。右下：認識結果

図 4　濃淡画像から机上シーンの認識

Figure 4　Desk scene recognition from gray image

このような場合、もし固定閾値で線を抽出すると、不要な線を含む多数の線が得られるが、重要であるがコントラストの小さい線が得られないことが多い(図 4 右上)。また、これだけ多くの線から物体を認識することも容易でない。

そこで、積み木の場合と同様に、画像から明確な特徴を抽出し、解釈を試み、それを手掛かりとして、つぎに明確な特徴を探して解釈するという過程を繰り返す。例えば、楕円が得られたら、その周辺のエッジを抽出し、電話機であることを確かめる。このようにすれば、信頼性のある線が最初に得られるので、間違いにくい。また、いったん認識されれば、その中にある不要な線を抽出することを避けることもできる。このようにして、同図右下のように、コントラストの低い線で構成される物体も認識できる。

この場合も、物体のどの部分から認識されるかわからないので、プログラムはやや複雑になる。さらに、学習による認識と比べると、対象物の種類ごとにプログラムを作らなければならないという欠点がある。利点は、人が考えて物体を定義しているので、その性能の限界がわかっていることである。

(4)　対象の属性が未知の場合

屋外風景などは、あらかじめどのような物体があるかわからない。たとえ、木があることが分かっても、その形や色は一定でない。このような場合に対する一つのアプローチ[10]を紹介する。

屋外の自然の風景を認識する場合を考える。そこには、木の幹、枝、葉、草、空、地面などが想定されるが、その属性は未知である。そこで、可能性のある属性のモデルを複数想定する（default model)。例えば、図 5 のように、木の幹の画素値（明度や色相）の分布を複数考える。それぞれに基づいて幹としての尤度を求める。

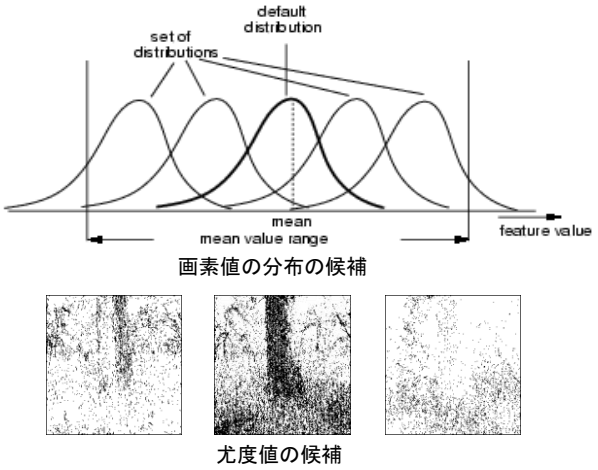


図 5　複数の画素値の分布と尤度

Figure 5　Multiple distribution and likelihood

この例では､ほぼ正しい分布を用いた結果(図 5 の真中)を含めた 3 種類が示されているが、実際はかなりの数の候補となる。この尤度に基づいて、幹の候補領域を求める。図 6 上段に、色相を用いて得られた複数の候補領域の例を示す。得られた領域から、幹らしいものを探す。これは、人が種々の環境でものを認識する場合､複数の解釈の中で、形のよいものや辻褄の合う解釈を採用することに相当する。

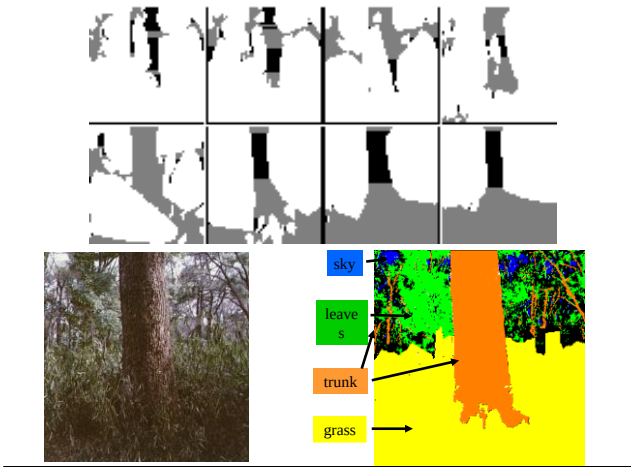


図 6　色相に基づく候補領域（上）と認識結果

Figure 6　Candidate region with hue(upper) and recognition result

色相以外の属性に関しても同様な処理を行い、最後にすべてを統合して幹を決定する。図 6 の下段に入力画像（カラー）と処理結果を示す

以上の処理は図 7 のように表され、原理的には膨大な量となる。しかし、同様なことを人が高速に行っているともいえよう。

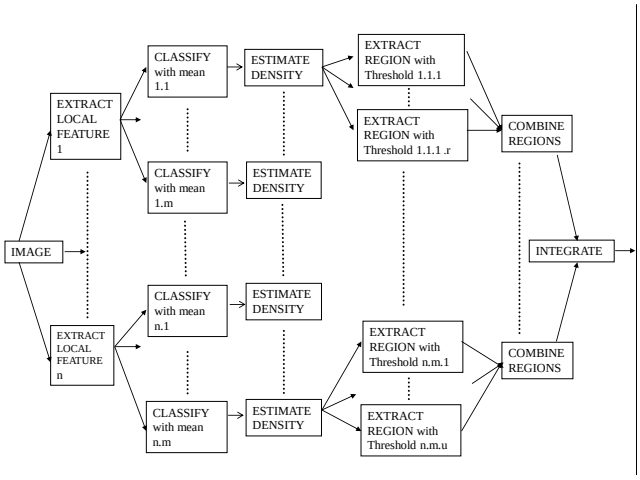


図 4　濃淡画像から机上シーンの認識
Figure 4　Desk scene recognition from gray image

この方式は、複数の属性の分布が既知としているが、これは正確である必要がなく、人が少数の例から学習したもので十分である。

6. おわりに

CV 研究はある程度飽和して、今後の発展はその応用を広げることという考え方も一理ある。確かに、医用、生物学、安全、エンターテインメントなど、CV の活躍が期待されている分野は多く、コンピュータの発展とともに、ますます盛んになろう。有意義な応用も CV の発展に役に立つので、手法の新規性にこだわることなく評価をし、論文としても公表していただきたい。

一方、すべての応用のために、複雑なプログラムを作ることは効率が悪い。そこで、OpenCV や Matlab などを利用することによって解決していることが多い。しかし、難しい問題に対しては、既存のプログラムをそのまま使えるとは限らない。透明性があり、改良が容易なプログラムの提供が望まれている。学会が主体的に取り組み、世界に通用するツールを提供していただきたい。

もう 1 つの解決法は学習を行わせることである。現在は多数の訓練例を与える方法が中心であることはすでに述べた。人が学習する場合も、訓練例は有用であるが、それに加えて対象を説明する記述を得ることにより、少ない訓練例でも汎用性のある能力を獲得できる。したがって、今後の学習も何らかでこのような機能を与えられることを期待したい。なお、本原稿をの締切間近に、はこだて未来大学の松原氏が、深層学習と記号推論の融合を目指すのが重要という認識を示したという報告[12]を読んだが、人工知能学会でも深層学習が注目を集めていることがわかる。

筆者らは、役に立つロボットビジョンのために、人がシステムとのインタラクションによって認識を助けるというインタラクティブ・ビジョンを提案した[11]。これは、認識時に人が助けるのであるが、学習時に人が助けることができれば都合がよい。現在も遺伝アルゴリズムのような一部の学習には人の介入があるが、CV には適用困難である。深層学習に人が簡単な概念を加えることが許されれば、より高度な能力を与えられるであろう。

人は、対象シーンに関する少ない事前知識から認識を行うことができる。著者は小さい問題ではあるが、人の能力への接近を試みてきた。達成した能力は今の水準にははるかにおよばず、プログラミングの労力が避けられないという欠点もがあるが､その考え方が参考になれば幸いである。

参考文献

1)　Pearson, K.: On lines and planes of closest fit to systems of point in space, Philosophical Magazine, Vol.2, pp.559-572 (1901).
2)　Essa, I. A. Sclaroff, S. and Pentland, A: A Unified Approach for Physical and Geometric Modeling for Graphics and Animation, Computer Graphics Forum, Vol.11, No.3, pp.129-138 (1992).
3)　神山泰典、松本俊昭、白井良明、島田伸敬、植田勝彦：ゴルフスイング診断のための画素値の時間ヒストグラムとグラフカットに基づく人体シルエット抽出, 電気学会論文誌, Vol.132, No.11, pp.1840-1846, (2012).
4)　Thien, B. N. and Shirai, Y.: Object Retrieving from Image Database, 信学技報, Vol.110, No.381, PRMU2010-168, pp.155-160 (2011).
5)　Shirai Y. and Tsuji S. : Extraction of the Line Drawing of 3-Dimensional Objects by Sequential Illumination from Several Directions, Proc. 2nd Int. Joint Conf. on Artificial Intelligence, pp.71-89 (1971).
6)　Shirai Y. and Suwa M.: Recognition of Polyhedrons with a Range Finder, Proc. of 2nd Int. Joint Conf. on Artificial Intelligence, pp.80-87 (1971).
7)　Shirai, Y.: A Context Sensitive Line Finder for Recognition of Polyhedra, Artificial Intelligence, Vol.4, No.2, pp.95-119 (1973).
8)　Sugihara, K. and Shirai, Y.: Range Data Understanding Guided by a Junction Dictionary, Proc. 5th Int. Joint Conf. on Artificial Intelligence, pp.706-707 (1977)　(情報処理学会論文では論文賞受賞).
9)　白井良明: 濃淡画像から複雑物体を認識する一手法, 情報処理, Vol.17, No.7, pp.611-617 (1976).
10)　Hild, M. and Shirai, Y.: Interpretation of Natural Scenes Using Multi-Parameter Default Models and Qualitative Constraints, Proc. ICCV'93 pp.497-501 (1993).
11)　Shimada, N. Miura, J and Shirai, Y.:Interactive Vision for Personal Service Robot, IPSJ Trans. on Computer Vision and Image Media, Vol.47 No.SIG15 CVIM16, pp.1-9 (2006)
12)　麻生英樹: 特別セッション「人工知能研究拠点の設立」, 人工知能, Vol.30, No.6, p.764 (2015)