4．寄稿

4.1 より柔軟な機械視覚をめざして

より柔軟な機械視覚をめざして

生体高次情報系　助教授　石川　博

Vision is the art of seeing things invisible.

— *Jonathan Swift (1667-1745)*

機械視覚の難しさ

私はコンピュータービジョン（機械視覚）という分野の研究をしています。それは一言で言って、いかにしてコンピューターに「見る」能力を与えるかということの研究です。例えば、ロボットに湯飲みをとって来させようとする時に、ロボットが湯飲みを見つけられなければなりません。我々は普通、目で見て見つけますから、ロボットにもそうさせようと、まず思います。ところが、これが大変難しいことがわかりまして、なかなかできません。一説によれば、人工知能の草創期の人々は、テレビカメラをコンピューターに繋げてそれを理解するプログラムを書くのはひと夏でできるだろうと思ったらしいのですが、そう簡単ではありませんでした。実際、最近ロボットの発展がめざましいかのようにいわれますが、視覚という面ではまだまだ実用になる場面は限られています。ロボットにこんなことができますというデモを最近よく目にします。例えば、皿洗いをしますというとき、ロボットはまるで人間のように皿を洗っているようですが、よく見ると、洗っている皿は、ピンクとか黄色とか、妙な色をしたものばかりであったりします。これは実は、背景にない色を皿だけに使うことで、皿を認識しているからなのです。

もっとも、計算機の能力が指数的に増大したおかげで、限定された環境下では、かなり実用的な機械視覚システムもできています。例えば文字認識とか、正面から見た顔の認識のように、対象となるものを限定でき、しかも、前もって個別に与えられた対象を見つけるというように、目的とするタスクも限定されている場合です。しかし、一般の画像から複雑な情報を引き出そうとすると、まだまだ実用には遠いのが現状です。例えば、食器を洗って棚にかたづけるロボットを考えます。食器というとかなり限定されているようですが、上に書いたように色で識別するとか、個別の食器を精密に指定してそれを探すのではなく、食器とい

うカテゴリーに属する物体を見つけて、さらにそれのどこをどうつかめばよいか立体構造を認識するとなると、はるかに難しくなります。アイロンをかけるロボットのために、任意の衣服を認識するのは、さらに困難です。人物写真で人間の輪郭だけをきれいになぞるという、人間になら簡単なことも、照明や背景がちょっと複雑になると、コンピューターにはできません。車を自動運転することがいまだ不可能であることの大きな理由の一つも、自動車を運転するときにまわりに認識する必要のある事象が非常に多様であることです。人間にとって視覚は最も自然な知覚であり、画像に映っている情景を理解することにあまりに慣れてしまっていますから、これを機械で実現するのがいかに難しいかということが、なかなか想像しにくいものです。例えば、画像は普通、長方形上の各点に色が指定されているものですが、これを色として表示せず、各点の明るさのグラフとして図１のように見せると、画像を画像として見る能力を使わずに理解しなければならないので、人間にとっても認識は難しくなります。本文末にある図３が元の画像です。それを見れば何が映っているのか明らかですが、これは人間に高度な視覚能力が備わっているからです。機械視覚とは、このようなデータを与えられて、それから意味的な情報を何とかして引き出そうという研究です。

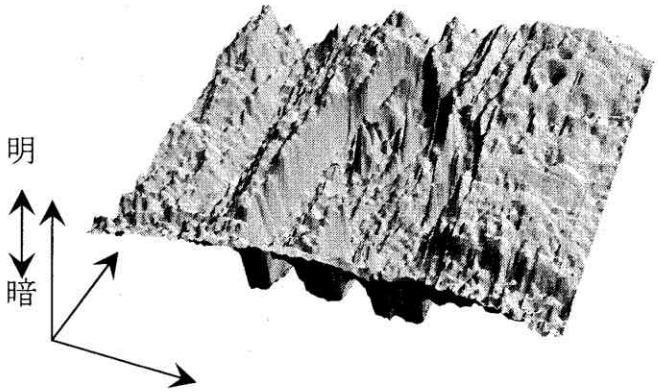


図１

なぜ難しいか

さて、なぜそんなに視覚が難しいかといいますと、これにはいろいろな理由があります。まず重要なのは、視覚が逆問題であるということです。コンピューターグラフィックスというものがあります。これは指定された３次元空間の状態を２次元の画像に帰する過程ですが、単純にいって物理的な光学系をシミュレートしてやればいいことになります。物理系のシミュレーションで、本質的に難しそうなことはなさそうです。（などというとその方面の方に怒られそうですが、あくまで相対的にという話です。）機械視覚はこの逆を考えるのですが、この逆の方が本質的に難しいのです。なぜならば、画像は必ずしも元の３次元情景の情報を

全部含んでいないからです。例えば、奥行きの情報は失われます。また、画像上である色に見える点があったとして、なぜその色に見えるかということは、その点に対応する視線上にある物体の反射率と、その表面の向いている方向と、そこにあたっている照明などが組合わさって決まるのですが、例えば表面に黒い模様があるのか、そこだけ面の向いている方向が違うのか、見えないところにある何かが影を落としているのかというようなことは、一般には判りません。この不定性のもっとも明らかな例は、ある情景の画像があったとして、その画像と、その画像の画像は見分けがつかないということです。このように、画像から情景に一意に変換することは原理的に不可能なのです。

もうひとつ、視覚が難しいということの、より本質的な理由があります。それは、画像から視覚で引き出したい情報は、ほとんどすべての画像について定義さえされていない種類の情報であるということです。画像とは何でしょうか。画像の存在を忘れるのは簡単です。画像を見る時、我々は「画像」を見ているのではなく、そこに映っている情景を見ているからです。肖像写真を見る時、そこに並んだ無数の微小な画素のことを考えることはありません。画像とは向こう側を覗くことのできる穴のようなもので、穴の存在はあまり意識しないのです。しかしそれこそが我々の視覚のなせるわざであり、機械にとっては、画像とはただ色のついた点の集まったものに過ぎないのです。色の集まりとしての画像、画素の色の組合せとしての可能な画像のほとんどすべては、我々が画像として普通に思い浮かべるものではなく、放送終了後のテレビのノイズのようなもので、いかなる情景を映したものでもありません。ところが視覚が画像から引き出したい情報は、画像についての情報ではなく、映っている情景についての情報ですから、ほとんどすべての画像について定義さえされないことになります。例えば画像の中で前景と背景を分離したいというとき、なにが前景であるかを定義することは、どんな色の画素がどのように並んでいるかといった、画像についての情報だけを使っては不可能です。そこには必ず、映っている情景についてのなんらかの情報が入っていなければならず、前景という概念が、例えば顔や文字という概念に比べてより一般的であるだ

図2

け、定義が難しいのです。人間なら、一枚の写真を見せられて、どこが前景であるか、すぐにいうことができますが、機械にこれを一般にさせることは、不可能です。我々が画像を見て認識する場合には、その画像内にない情報をかなり使っています。図２の画像は、一度でも見たことがあればすぐに何が映っているかわかりますが、見たことがないと、普通しばらくの間は黒いしみがあるだけにしか見えません。ところがこれを見つづけていると、突然、なにか３次元の情景が見えてくるはずです。この画像を理解するには、この世界についてのかなりの知識が必要であることは明らかです。これは極端な例ですが、人間の脳は、画像情報にあてはめる、かなり強力なモデルをもっていること、つまり世の中にはどんなものがあるかという大量の知識を持っているということを示唆しています。この知識とは別に犬の種類を知っているということではありません。例えば写真にノイズが入っていると、我々にはノイズだと判ります。これは自然界にはノイズのようなものが見えることはないということを、あるいはその確率は非常に低いということを、視覚系が知っているからです。またこれは今まで見た画像を単純にすべて覚えているということでもありません。初めて見たものでも、我々はかなりうまく認識することができます。図２を一度も見たことがない人が、ちょうどあのような情景を見たことがあるということはまずないですが、それでもこの図を３次元の情景として認識できるのです。

柔軟なモデルの必要性

このように、「見えるものについての予想」を持っていることが視覚には必要です。文字認識が実用になりやすかったのは、結局、初めは数字のみ、次はカタカナという風に、限られたものを画像の中から見つけ出すという形式に、問題を定義できたからなのです。そこで、より一般的な機械視覚を実現しようと思うと、この「見えるものについての予想」をいかにして表現するか、またそれをどうやって機械に与えて蓄えるか、そしてそれを使っていかに画像から情報を引き出すか、ということが大きな問題となってきます。

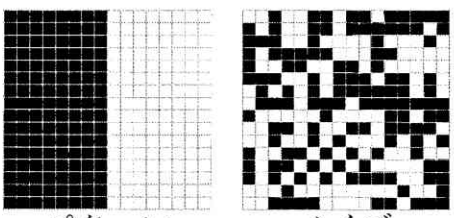
文字認識ならば、すべての文字のデータを画像として持っていて、画像中に探せばいいと考えられます。正面から見た顔の認識もそれに似て、データベースにある顔を画像中に直接探すことで実現しています。ちょっと見る角度の違う顔の認識も、コンピューターの能力が上がるに従って、いろいろな角度から見た顔を蓄え、探すことによって可能になってきています。しかし、このような方法では、もっと一般的な物体、例えば部屋の中にあるすべての物体を認識することはできません。ここで「認識する」とは、その物体が何であるかを分類するというより、それが１つの物体であると判って、例えば背景から分離することができるということです。上に述べた食器洗いやアイロンかけの例で言えば、食器１つ１つを物体として認識できたり、衣服を衣服として見つけられたりするということが第一歩になるからで

す。この、画像を意味のある部分に分割する、例えば図２なら犬と背景に分割する問題を、領域分割問題といいますが、コンピュータービジョンの難しさを端的にあらわす問題としてよく研究されています。一見簡単なようですが、領域分割は、画像を分割する問題でありながら画像についてだけ考えていては定義すらできない、非常に難しい問題です。

さて、私見によれば、「見えるものについての予想」をいかにして表現するか、またそれをどうやって機械に与えて蓄えるか、そしてそれを使っていかに画像から情報を引き出すかという問題を解決するためには、ただ画像のデータベースを持っていることよりも柔軟な、画像に映っている対象の持つ構造をとらえるモデルが必要であると思われます。ここで構造をとらえるとは、次のようなことです。画像のデータとしての表現は、普通、長方形に並んだ画素のそれぞれがどんな色であるかを指定することによってなされます。この表現方法は一般的で、どんな画像でも表現できるのですが、上に述べたとおり、逆に言えば、情景を映している写真のような普通の画像と、放送終了後のテレビのようなホワイトノイズを区別しません。この方法で表されうる画像のほとんどすべては、人間が見たら区別がつかないホワイトノイズなのです。ホワイトノイズには構造がありません。人間にはどれも同じに見えるのはそのためです。

この、構造あるいはパターンを持つ画像とノイズとを区別するものは何かということを考えます。例えば 100×100 画素の画像ならば、これは一万次元の単なるベクトルではなく、どの画素とどの画素が隣同士であるかといった構造を持ちます。画素をランダムに並べ換えれば、パターンのある画像でもノイズになってしまいますから、画像にパターンがあるということは、この画素の並び方といった幾何的構造について相対的に決まることです。ですから、この構造は、画像を扱ういかなるシステムにとっても一番基本的であり非常に重要です。しかし従来、これに対する処理が実際にコンピューターにプログラムされる段になると、これらの構造はその場限りのデータの扱い方をするプログラムの中に隠されてしまうことが多かったのです。

私は、構造をとらえてモデルを表現するデータ表現方法の開発が、より柔軟で一般的な機械視覚を実現するための鍵であると思います。グラフという、頂点とそれを結ぶ辺とからなる抽象的データ構造があります。画像を、例えば画素という情報を持つ頂点と、それらの間



パターン　ノイズ

の相対的位置関係を表す辺とからなるグラフとして抽象化すれば、どれだけの構造を扱っているのか、その構造はどのようなものであるのか、またその構造に対してどのような操作が可能であるのかが明確になり、理論上判然とするだけでなく、機械による自動的な扱いを、より柔軟なレベルで可能にします。私はこれまで、機械視覚の分野を中心に、画像その他の高次元データに使うことのできる、グラフを用いたアルゴリズムの研究をしてきました。そのなかで、このように情報処理過程を抽象化することの、特に自動化の上での重要性を痛感しました。

グラフは単純にして豊富な構造を表現可能なデータ構造ですが、それでも表現できない構造は画像には多数含まれます。たとえば、一本の直線が画像中を走っていれば、非常に目につきますが、グラフでこれをうまく表す方法はありません。これからもわかるように、一つのデータ構造ですべてを表現することは難しく、データに自動的かつ柔軟に対応するには、それを表現するデータ構造自体を操作する機構が必要であると思われます。今後の研究では、これをもう一段抽象化し、データ構造一般を自動的に取り扱う方法を探ることにより、機械視覚の柔軟性を向上させることを目指したいと思います。より柔軟で一般的なデータ表現方法を開発し、データとそれを特徴付ける構造をできるだけ抽象化し、データを扱う手続をも抽象化することが目標です。

ここには、計算機科学における成功例の 1 つであるプログラム言語の理論との類推があります。プログラム言語理論では、記号列の集合としてのプログラム言語の範囲を決めるシンタックスと、プログラム言語による記号的表現と計算機の動作を結びつけるセマンティクスを、抽象的にモデル化することにより、記号列を計算機の動作に自動的に変換するコンパイラの理論を作り上げ、大きな成功を収めています。自動的にコンパイラを生成するコンパイラコンパイラなるものまであります。これと同様に、機械に構造を扱わせるには「構造というもの」を抽象化してやる必要があると思うのです。

その先

画像のような情報の持つ構造を捉えたデータ表現は、コンピュータービジョンシステムの出力は何であるべきかという問題にも関連します。画像から何らかの意味的な情報を引き出すのがコンピュータービジョンの目的なのですが、引き出された情報をどう表現すればよいのでしょうか。言語による表現がまず思い浮かびますが、どうやって画像から言語に結びつけるべきでしょうか。画像情報と言語情報には相補的な関係があります。百聞は一見に如かずというように、画像には言語で表せない内容を伝える力がありますが、逆にその諺を、画像を使って明確に伝えることも困難です。これから解るように、両者の表現する情報には何か本質的な違いがあります。言語とは異なる情報の表現が何か必要だと思われる理由です。

人間の脳の中にある情報はこの二つだけでなく、例えば体を動かした時の感覚とか、すべて何らかの形で脳内に表現されているはずで、ロボットに同じことをさせようとすれば、やはり柔軟で一般的なデータ表現方法は何であるかという問題が生じます。そして、これらの表現を互いにいかに結び付けられるかという問題は、視覚に限らない人工知能というものを考える上において、避けては通れない問題です。知覚レベルと記号レベルをいかに結びつけるかということは、古典的な記号的人工知能の問題点として知られ「接地問題」と呼ばれるものと関連します。記号とは任意のものであって、記号操作による人工知能では人間の介在なしに「ゴリラ」という記号を実際のゴリラと関係づけられないという批判です。画像のような構造を持つ情報の研究を通じて、この問題にも何らかの光を少しでも当てられれば幸せだと思っています。

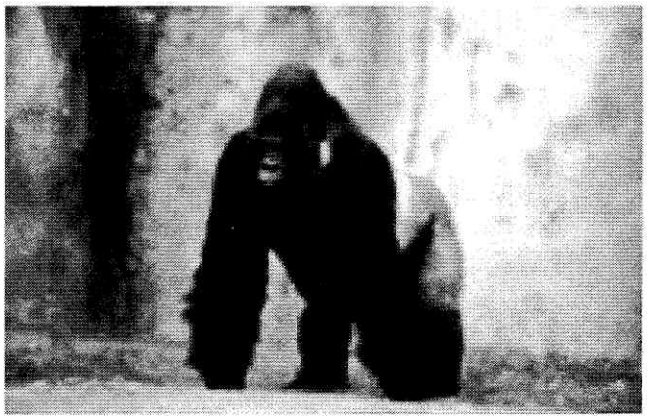


図3