ISSN 1883-1966(Print)
ISSN 1884-0817(Online)
国立情報学研究所ニュース

NII Today

National Institute of Informatics News

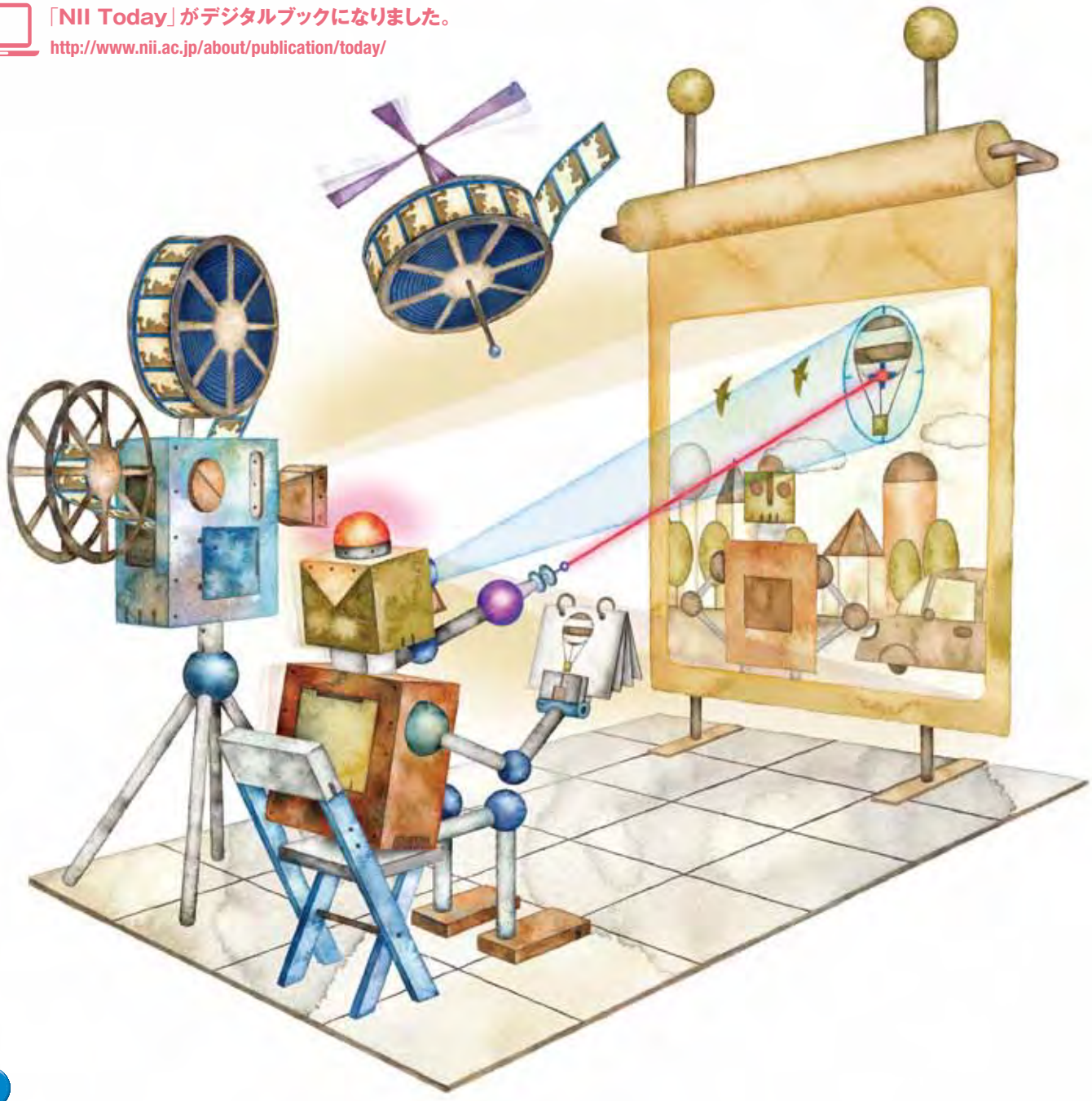
67
Mar. 2015

Feature

映像

情報技術が生み出す新潮流

「NII Today」がデジタルブックになりました。
http://www.nii.ac.jp/about/publication/today/



国立情報学研究所
National Institute of Informatics

NII Interview

テレビは社会の動きを感じ取るセンサーだ

テレビ、ラジオ、新聞などのメディアは社会を映す鏡。トレンドを感じ取るセンサーでもある。映像の意味を解析するシステムの研究に取り組む国立情報学研究所の佐藤真一教授と、テレビ番組の放送内容を記述したTVメタデータの提供、調査、分析をする株式会社エム・データ※の小口日出彦取締役に、センサーとしてのテレビ利用の現状と未来を話し合ってもらった。

※株式会社エム・データ　http://mdata.tv

辻村　まずはどんなデータを取り出し、利用しているか教えてください。

小口　私たちの会社はテレビに、いつ何がどのように映っていたかという情報を集めています。テレビは映像、音声、テキストを総合したメディアです。それらの情報を扱いやすくするため、全部文字に起こしています。

　それは本の「書誌データ」みたいなものです。タイトル、著者、ページ数、ジャンルなどのデータを基に私たちは目当ての本を探しますよね。書誌データはデータの集まりである本に関するデータ。「メタデータ」と呼ばれるものの1つです。

　テレビの映像にも必ずメタデータが付いている。それを文字にしてデータベースを作る。例えば「ゴジラ」と入力すると、過去１年間にゴジラが、いつ、どんなふうに放送されたかがわかる。あるものが頻繁にテレビに出るというのは、世の中のトレンドを表象する情報発信です。視聴者がそれを見ると、その行動は情報を送る側にフィードバックされ、さらに出る場面が増える。テレビはそんな世の中の情報の流れをとらえるセンサーであるとも言えます。

佐藤真一
SATOH Shin'ichi
国立情報学研究所
コンテンツ科学研究系 教授・主幹

辻村　佐藤さんはどんな研究をしているのですか。

佐藤　20年ほど前、画像や映像の検索技術の研究を始めました。色の分布を基に、数千から数万の画像の中から、似たものを検索する技術が出てきたころです。ただ当時は動物だけとか建物だけとか、限られた画像の中から検索するだけでした。さまざまな映像が流れるテレビから、欲しい情報を取れれば、実用的な技術になるのでは、と考えました。

　まずテレビ放送の録画システムを作り、これまでに40万時間分の映像を蓄積しました。映像から研究の邪魔になるCMを抜き取る技術も開発しました。抜いたCMのどれが同じかも判定できる。すると、あるCMがどの曜日のどの時間帯にどれだけ放送されたかがわかり、そのパターンを分析すると、企業の消費者戦略が見えてきました。リーマンショックの後だったせいか高額商品のCMがほとんどないとか、景気が上向くと高級車のCMが増えたとか、これは社会のセンサーとして使えると思いました。

小口　テレビは情報源として私たちの生活に影響を与えています。マーケティングの人たちは、その情報がいつどんなふうにテレビで語られたのか知りたい。それに役立つ仕事ができるのではないかというのがエム・データという会社の始まりでした。

小口日出彦
KOGUCHI Hidehiko
株式会社エム・データ 取締役

　一方で、世界中の情報が流れる交差点のようなところに網を張り、俯瞰してみたいという夢ももっています。例えば国内の政治問題が国外で報じられる際の違いを相対的に扱えるようになれば、物事を中立的な視点から見て判断したり選択の幅を広げたりできるのではないかと考えています。

　情報にはどうしてもバイアスがかかる。どんなバイアスかが見えれば、人々に与える影響は大きい。そんな「メディアのメディア」のようなものができれば面白いでしょう。

佐藤　実際にどうやってメタデータを取っているのですか。

小口　専門スタッフ約80人が輪番制で、24時間365日、テレビを見ながら情報を入力しています。自動化できるといいのですが、現状では機械による文字化はどうしても数％程度は間違える。それを人間が直すくらいなら初めから人間が識別し判読した方が圧倒的に速いのです。

佐藤　私たちは「人間が映像を見てメタデータを付けたらいいのでは」という批判を浴びます。映像が大量なら全部見るのは大変だと反論しているのですが、エム・データはそれをしておられるので困りますね。（笑）

小口　いえ、ご研究には意味があると思います。例えば自動車のF1レースでコースに掲げられた企業のロゴが何秒間、画面のどこに映ったかという情報は企業にとって重要です。それを自動的に分析できるなら、人間よりもずっと効率がよい。

佐藤　私たちは「物体検索」の研究を続けていて、これはロゴ検出に最適なんです。世界的にも競争力のある検索精度をもつ技術で、かなり検出できるようになりつつあります。

　先ほどの話の続きですが、ニュースの要約を作るなどメタデータを取り出すとき、人間の主観が入ることはないですか。

小口　情緒的な判断が入らないように入力インターフェースを工夫しています。できるだけ主観の入り込む余地がない対象を抽出するので誰が入力してもデータの揺れは少ない。

佐藤　映像から情緒的な情報を取り出すことが重要になる場合もあるのでは。機械で立ち入るのは難しい領域ですけれど。

小口　直接把握するのは難しいのですが「露出の頻度や時間」を分析すると、例えばある物事がセンセーショナルに扱われているような現象をとらえることができます。

佐藤　ウエアラブルなセンサーを使って、テレビを見る人がどんな刺激を受けているか把握するというのはどうでしょうか。

小口　すぐの事業化は難しいですが、やってみたいですね。テレビを見ている人がリアルタイムでツイッターに何を書き込み、どう共感が広がっているかというのはモニタリングの対象に入れつつある。人間の生活に関わる記録をテレビのメタデータと併せて取り扱おうというアプローチも始めています。

インタビュアーの一言

対談は話題を変えつつ盛り上がり、予定時間を超えて続いた。印象に残ったのは「技術力は高いが、それを転がすアイデアがない」「基礎研究を粘り強く支える体制がない」という日本の課題だ。そのためにコンピュータ産業も、画像認識技術も米国に先行された。研究開発の進め方を見直す必要がある。

辻村達哉
TSUJIMURA Tatsuya
共同通信社 編集委員 論説委員

1984年、東北大理学部物理第2学科を卒業後、共同通信社入社。大阪支社社会部、大津支局、札幌支社、釧路支局を経て本社ラジオ・テレビ局報道部、科学部。2005年から編集委員室兼務。06年にメディア局メディア編集部。07年から編集委員兼論説委員、10年に秋田支局長、13年9月から現職。

NII Special 1

実時間で 3次元形状モデルを生成する

安価になったRGB-Dカメラを活用

NIIの杉本晃宏教授は、画像に加えて距離の情報を取得できるRGB-Dカメラで取り込んだデータに基づき、3次元形状モデルを効率よく生成する手法を研究中である。効率がよい3次元形状の表現手法を独自に考案し、従来手法よりも劇的にメモリ消費量を減らした。人間の顔、表情を3次元形状モデルとして生成し、スマートフォンからSNSに送るといった応用も視野に入れている。

リアルタイム 3次元形状データを 手軽に扱える時代がくる

杉本晃宏教授がRGB-Dカメラで取り込んだ画像データから3次元形状モデルを生成する研究を開始した1つのきっかけは、RGB-Dカメラが圧倒的に安価になったことだった。RGB-Dカメラとは、カラー画像（RGB）だけでなく、対象までの距離（D）を取得できるセンサーである。距離を測るセンサーとしては、最近まで高価なレンジファインダーが主流だった。数年前、その状況が変わった。2010年に米Microsoftがゲーム機用の周辺機器として発売した「Kinect」を筆頭に、民生用の安価なRGB-Dカメラが入手可能となってきたのだ。

「そこにKinectがあったから、この研究をやりたくなった――といったら言いすぎかな。でも、安くてより多くの人が使える、そういう新しいものをやりたいと思いました」と杉本教授は話す。

杉本晃宏
SUGIMOTO Akihiro
国立情報学研究所
コンテンツ科学研究系 教授
総合研究大学院大学 複合科学研究科
情報学専攻 教授

RGB-Dカメラは今後ますます安価になり小型化され、スマートフォンに搭載されるなどの形で普及が進むと予想されている。しかもプロセッサの高性能化や、画像処理のためのGPU（グラフィックプロセッサ）の活用も進む。そうなれば、従来のカメラのように2次元の動画像データだけでなく、3次元形状を実時間で取り込むことも可能となってくる。例えば、スマートフォンに搭載したRGB-Dカメラでデータを取り込んでリアルタイムに3D形状モデルを生成し、その場でソーシャルネットワークにポストするといった応用も視野に入ってくるのだ。

実際、すでにKinectとノートPCのGPUの活用により3次元形状のリアルタイム処理を実現できている。近い将来は、スマートフォンでも同様の処理が可能となるはずだ。

3次元形状を、 2次元データを使って操る

杉本教授の研究で特に注目したい点は、3次元形状の表現手法を新たに考案したことだ。これによりメモリ消費量を劇的に減らすことができ、より大規模かつリアルタイムな3次元形状のモデル化が可能となった。メモリ消費が小さいことは、データを処理する上でも、通信で送る上でも有利である。

3次元形状の表現で主流となっている手法は、3次元、つまり3つの座標軸で表現した空間を「ボクセル（voxel）」と呼ぶ最小単位で区切り、デジタル情報として扱うやり方だ。このやり方は無駄が多い。何も存在しない空間に対しても多くのボクセルを割り当てる必要があり、膨大なメモリを消費するからだ。

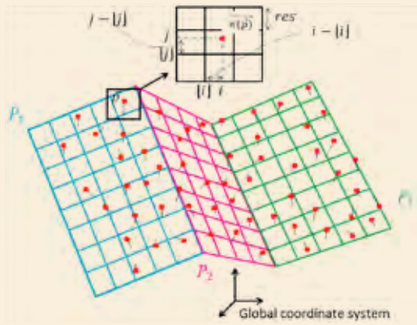
一方、杉本教授の手法は、2次元の平面を使って3次元の空間を表現する。3次元の各点を平面上の点に対応させ、3次元の点が対応する平面上の点からどれだけ「ずれ」があるかの情報を画像としてもたせることで、3次元形状を表現するのである。膨大な数の3次元のボクセルを扱うのではなく、いくつかの2次元の平面のデータを扱えばよい。いわば、3次元を2次元のデータで操るのである。

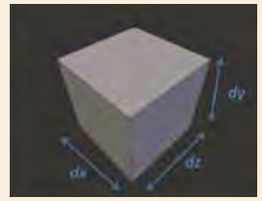
この手法により、メモリの消費量を劇的に減らすことができた。モデル化したい「もの」のサイズが大きな場合や、スケールアップ（拡大）をしたい場合、従来手法ではすぐにメモリを消費し尽くしてしまい、限られた範囲でしかモデル化できなかった。それが杉本教授の手法を使うことで、より大きなサイズ、大きな縮尺の「もの」の3次元形状をモデル化できるようになった。

「どういう環境かにもよりますが、従来手法だと、5m四方の部屋をモデル化するのが精一杯でしたが、私たちの技術を使えば部屋の外側の空間もリアルタイムでモデル化できます」（杉本教授）

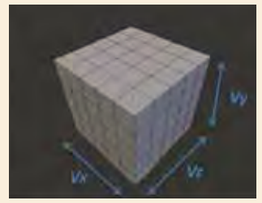
しかも、2次元のデータで表現する手法は、画像処理の分野で長年研究されてきたさまざまなテクニックを応用できる。例えば、普及している画像圧縮技術を応用することも可能だ。

3次元形状を2次元の平面データとして操る手法。
扱うデータ量が圧倒的に少なくてすむ。





dx = dy = dz = 3 [meters]



Vx = Vy = Vz = {32, 64, 128 ,256, 512} [voxels]

「理論と応用を全部やろう」と、 研究テーマを設定

杉本教授は、理論研究と応用研究を並列に進めるスタイルを採ってきた。

杉本教授が長年取り組んでいる研究テーマとして「離散幾何」がある。デジタル化、つまり離散化に伴う誤差は、どんなに解像度が向上しても必ず存在する。その理論的な限界を知るための研究だ。デジタル画像が高解像度になってくると、デジタル画像の解像度には限界があることを忘れがちだが、「元の画素の精度を忘れて、復元する精度だけを追求しても意味がない」と杉本教授は言う。

このような理論の研究を手がける一方で、RGB-Dカメラを使った3次元形状の復元といった応用研究にも取り組む。杉本教授は「誤解を恐れずにいうなら、理論に重心をおく研究者と、応用に重心をおく研究者は、旧来、互いに批判しあう傾向があった。自分としては、『じゃあ全部やろう』と意識している」と明かす。2次元データで3次元形状を表現する独特の手法も、このような研究スタイルから生まれてきたアイデアなのかもしれない。

顔や表情の再現から、 街全体のデータまで

杉本教授が3次元形状のモデル化として取り組んでいる対象の1つに、「顔」がある。人間の顔の3次元形状をリアルタイムで復元できれば、微妙な表情の変化をより正確に伝えることも可能となる。「例えば、遠隔地にいる高齢者の精神面のケアをする際には、表情が大事になります」と杉本教授。このようなデリケートな目的のための通信では、人間の表情をより正確に表現するデータを伝送することが重要になってくるのだ。

今後の展望としては、都市の3次元形状のデータを扱う動きがある。「東京オリンピックが開催される5年後には、東京の街全体の3次元モデルはできていると思います」と杉本教授は話す。こうした巨大な3次元データの集積が進む一方で、RGB-Dカメラで手軽に3次元形状を取り込めるようにもなる。多種多様な3次元データを普通の人々が日常的に使うようになる日は、そう遠くなさそうだ。

（取材・文＝星 暁雄）

「グラフ信号処理」で進化する臨場感のあるコミュニケーション

画像・映像の圧縮・補間・ノイズ除去効率を上げる新手法とは？

「グラフ信号処理」技術は、ソーシャルネットワークやセンサーネットワークからの信号を適切・効率的に分析するのに利用される比較的新しい処理技術。その技術を画像や映像に適用する研究の先駆者である南カリフォルニア大学のアントニオ・オルテガ教授と、長年ともに共同研究を続けているNIIのチョン・ジーン准教授に、グラフ信号処理技術が「臨場感のあるコミュニケーション」および画像・映像の世界にどんな革新をもたらすのかを聞いた。

アントニオ オルテガ
Antonio Ortega
国立情報学研究所 客員教授
南カリフォルニア大学 教授

チョン ジーン
Cheung Gene
国立情報学研究所
コンテンツ科学研究系 准教授
総合研究大学院大学 複合科学研究科
情報学専攻 准教授

「点」と「線」で信号の構造を明らかにする

――「グラフ信号処理」を簡単にご説明ください。

オルテガ　グラフはノード（点）とエッジ（線または辺）で情報の構造を表現する方法です。わかりやすい例では、FacebookなどのSNS参加者をノード、参加者相互のつながりをエッジとしたグラフをイメージしてみるとよいでしょう。

個々のノードは、性別や年収、好みの音楽、興味のある分野など、多様な属性をもっています。それらの情報はノード上の「信号」とみなせます。各ノードは、主に「友達関係」で結ばれますが、さらに、「同程度の年収」「好みの音楽」「興味分野」などが共通するノードと結ぶことも可能です。また属性の共通性や類似性の度合いにより、エッジにウエイト（重み付け）を与えることができます。SNSをマーケティングなどの目的で分析したい時には、目的に合わせてノードとエッジの構造とエッジウエイトを適切に設計することで、効率的・効果的に市場動向や顧客属性などを把握できるのです。

気象センサーのネットワークでも同様です。地域にメッシュ状に配置された気象センサーは、隣り合うセンサー同士でネットワークを構成し、温度や湿度、降雨量などのデータを取得、計測値やセンサー間の距離などにより、各エッジに重み付けを行って、地域内の気象変化を把握します。このように、ノードとエッジからなるグラフを作成して、そこに発生する信号間の関係性を調べることでデータを効率的に分析しようというのが、グラフ信号処理の基本です。

ジーン　そして、グラフ信号処理の技法を画像処理に応用しようというのが私たちの共同研究です。従来の画像や映像の信号処理は、音声なら時間間隔で等分し、画像ならピクセルを等間隔でサンプリングし、映像なら加えて等間隔のフレームに分割して処理してきました。いわば均一に整った「レギュラー」な構造を前提にした処理です。一方、グラフ信号処理は、「イレギュラー」な構造をもった信号を対象にできる。必要に応じて適切なグラフを自由に作成し、信号間の関係性を解釈することで、従来にない効果が得られるのです。

図1　奥行き画像の変換の例

メッシュ上に整列しているピクセルの中で、類似性が高いもの（この場合はカメラからの距離が同程度のもの）同士のエッジウエイトは1、違いが大きいもの（カメラとの距離が遠いもの）の間のエッジウエイトは0とする。するとグラフの線の一部は切れることになり、2つの部分（前景と背景の画像）に分割できる。

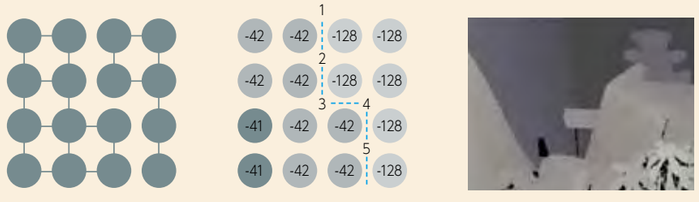


図2　グラフ信号処理によるノイズ除去の例

類似する部分の平均をとってノイズを除去する。グラフを設計して適切なエッジウエイトを付与することで、画像がぼやけずに鮮明な状態でノイズを除去できる。

臨場感のあるコミュニケーションに不可欠な圧縮・補間・ノイズ除去効果を改善

―― 臨場感のあるコミュニケーション研究にグラフ信号処理はどう役立ちますか？

ジーン　臨場感のあるコミュニケーション研究は、例えば遠く離れた場所にいる人とテレビ会議をする場合、ディスプレイに対面している人が映像の中の人の肩越しに後ろを覗きこめば背後にあるものが見えるような、現実とほとんど差のない体験の実現を目指しています。これを発展させれば、例えばスポーツの映像観戦でも、観客席を自由に移動するようにして、自分が見たい角度から観戦できるようになるはずです。

その実現には、多数のカメラで撮影した映像が必要で、見る人の「視線と視点」の変化を予測しながら必要な画像を即座に合成しなければなりません。データ量が膨大になるので、効率よい伝送や処理のためにデータ圧縮は不可欠ですし、データが抜け落ちた部分を自然に見えるように補う補間も大切です。また画像のノイズを排除する必要もある。グラフ信号処理は、「圧縮」「補間」「ノイズ除去」のいずれにも効果が期待できます。

例えば、画像のカメラに近い被写体と背景を切り分けたいとき、類似度が高いノード（ピクセル）間のエッジは高いウエイトにし、大きな違いがあるノード間のエッジは低いウエイトにしてグラフを作成します。グラフフーリエ変換をすると、最低周波数の交流成分のプラス・マイナス記号により、前景と背景を効率的に切り分けることができます。

画像圧縮にこの方法を応用すれば、1つのブロックに対して、前述した画像の構造を表すグラフを用いてフーリエ変換を定義すると、低周波数成分しかない表現が得られます。そのスパース（疎）な表現により、圧縮率を上げることができるのです。臨場感映像で映像合成に利用する奥行き画像で試したところ、30～40％の高圧縮率を実証しました。

また、画像の一部のデータが欠落した場合でも、周辺のピクセル間のエッジのウエイト変化に合わせて、スムーズに見えるようにデータを挿入して補間できる。ノイズ除去は、画像中の類似した部分の平均値をとってはめ込みますが、エッジウエイトを適切にとることにより、被写体をぼやかすことなく、ノイズだけを取り去ることができます。

画像・映像領域での「グラフ信号処理」の今後

――今後の抱負をお聞かせください。

オルテガ　グラフ信号処理の画像・映像への応用は研究段階ですが、特に圧縮については企業の注目度が高く、GoogleやMicrosoftも研究しており、まずは企業独自技術として3～4年後には利用が始まり、その後、標準化に向かうかもしれません。他の応用例も、いま続々と出てきている最中です。今後も新たな応用を発見していくのが楽しみですね。

ジーン　画像や映像処理にまつわる古くからの課題に対して、グラフ信号処理は新しい視点から解答をもたらしてくれるところが面白い。オルテガ教授と手を携えて研究を進めることで、臨場感のあるコミュニケーションの課題になっている圧縮や補間などすべてに応用できるようなすごくパワフルなツールとして、グラフ信号処理技術を活用できるようにしたいと思っています。

（取材・文＝土肥正弘）

映像・画像解析と自然言語処理の
タッグで生まれる新しい価値とは?

映像と言語の協調で見えた意味解析の地平

That's **Collaboration**

映像・画像解析の分野に、テキストの解析を行ってきた自然言語処理研究者が参入しつつある。これまで異なったアプローチをとってきた両研究者の共通点と相違点、そして目指すべき方向は何か。自然言語処理の研究に取り組む東北大学の乾健太郎教授、NII宮尾祐介准教授と、NIIで映像・画像解析等の研究を手掛ける佐藤真一教授に、意味解析の過去、現在、そして未来について両者の立場から語り合って頂いた。

映像・画像を機械が解析する難しさ

乾　まずは映像・画像解析における研究の変遷についてお聞かせください。

佐藤　コンピュータが登場した頃から映像・画像の意味解析に対する要望は大きく、1960年代にはすでに研究が行われていました。人間が簡単にできることから、当時は機械でも容易に行えるだろうと考えられており、人間の視認性をプログラム化しようとしたり、画像に対してルールを定義したりすることで解析が可能とされていたのです。しかし、ほどなくして映像・画像解析はとてつもなく難しいことがわかりました。とてもシンプルな画像でも、途方に暮れるほど多くのルールを定義しないと、意味を解析できなかったのです。そうしたことから80年代半ばには、一度、多くの研究者がこの分野から手を引いたという経緯もありました。

宮尾　その後、どのような転機が訪れたのでしょうか。

佐藤　変化が訪れたのは90年代で、その成果の一例が顔認識です。「どこが目で、どこが口なのかをコンピュータに教え込む」、といった従来の手法をやめ、顔の学習データを大量に収集、とにかく「これが顔だ」と、機械学習によってコンピュータに覚え込ませたのです。そうしたビッグデータ的なアプローチが功を奏し、2000年代におけるデジタルカメラの顔認識機能の礎ができました。その後、森羅万象を認識させるためには、膨大な学習データを整えれば対処できると考え、現在では約2万2,000の概念に基づいて、約1,400万画像を収容するに至っている巨大なデータベース「ImageNet」が2010年に構築され、さらに2012年「ディープラーニング」が登場してきたことで、映像・画像解析は質的向上を遂げました。しかし、解析にあたってその基盤となる、“概念”をいかに選んでいくかが課題として挙げられています。

乾 健太郎
INUI Kentaro
国立情報学研究所 客員教授
東北大学 大学院情報科学研究科 教授

佐藤真一
SATOH Shin'ichi
国立情報学研究所
コンテンツ科学研究系 教授・主幹

乾　なるほど。詳しくお聞かせください。

佐藤　画像解析にあたっては、「画像とテキストによるシンボルとの対応づけを行えばよい」と考えられていたのですが、どの概念を選択すればいいのかが難しい。概念数も1万を超えると概念間で親子関係が出てきます。「乗り物」という概念の下には「自動車」や「飛行機」があり、そうした定義もきちんと行っていかなければなりません。また、「鷲が飛んでいる画像」と「ジェット機が飛んでいる画像」など、意味の関連性と見た目の関連性の一致・不一致についても正しく定義する必要があります。そうしたさまざまな概念の定義を整備し、画像認識の精度を向上させるための研究が進められている段階です。その一方で、いまだにコンピュータは未知の画像を与えられて「これは何ですか？」と聞かれるのが最も不得意なのですね。例えば画像だけを見て、それがイヌなのかネコなのかを答えることは難しいのですが、イヌという前提を与えたうえで、「この犬種は何ですか」と聞いた場合にはかなり高い精度で正解を出すことができます。そこに何か活路が見い出せるのではないか、と試行錯誤しています。

乾　そこはディープラーニングの活用によっても難しいのでしょうか。

佐藤　まだまだですね。あるデータセットでは、かなりの精度で画像を解析できるようになっています。ところが、どのような学習により認識できたのかを分析してみると、例えば「家」をうまく認識したのは、実は家の形状自体ではなく、周りの植え込みなどを見て判断していることがわかりました。家の形状は千差万別であり、コンピュータ側からすれば、むしろ植え込みを見たほうが精度の高い答えが出せる、というわけです。つまり、うまくいっているように見えても、本当の意味での解析には至っていないのです。

映像・画像解析と同じ歴史を辿った自然言語処理

宮尾　非常に単純な自然言語側からのアイデアなのですが、言語モデルや、コンテキスト情報の活用といったアプローチはとられていないのでしょうか。

佐藤　コンテキスト情報の活用も方策の1つとして挙げられていますが、それ自体も体系的に整備してつくっていかなければなりません。そうなると、画像解析の研究に留まらなくなってしまう。また、あくまでも画像解析における研究者のホーリー・グレイル、つまり求める“聖杯”は、画像だけを学習データとして与えて、未知の画像から解析結果を出すこと、という意識が研究者の中にあることも否めません。

宮尾祐介
MIYAO Yusuke
国立情報学研究所
コンテンツ科学研究系 准教授
総合研究大学院大学 複合科学研究科
情報学専攻 准教授

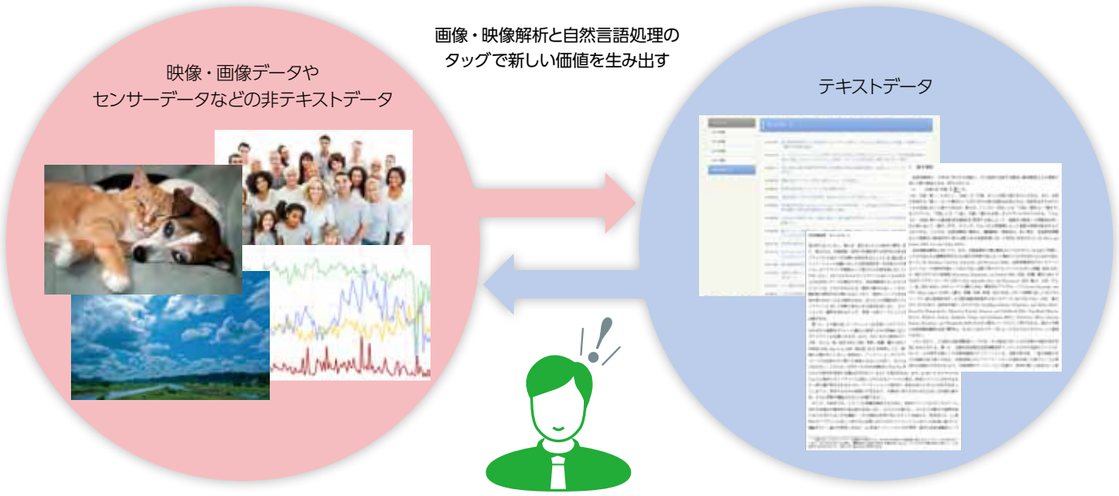
宮尾　それは驚きですね。自然言語側からすれば、使える情報はなんでも使った方がよい、と考えるのですが。

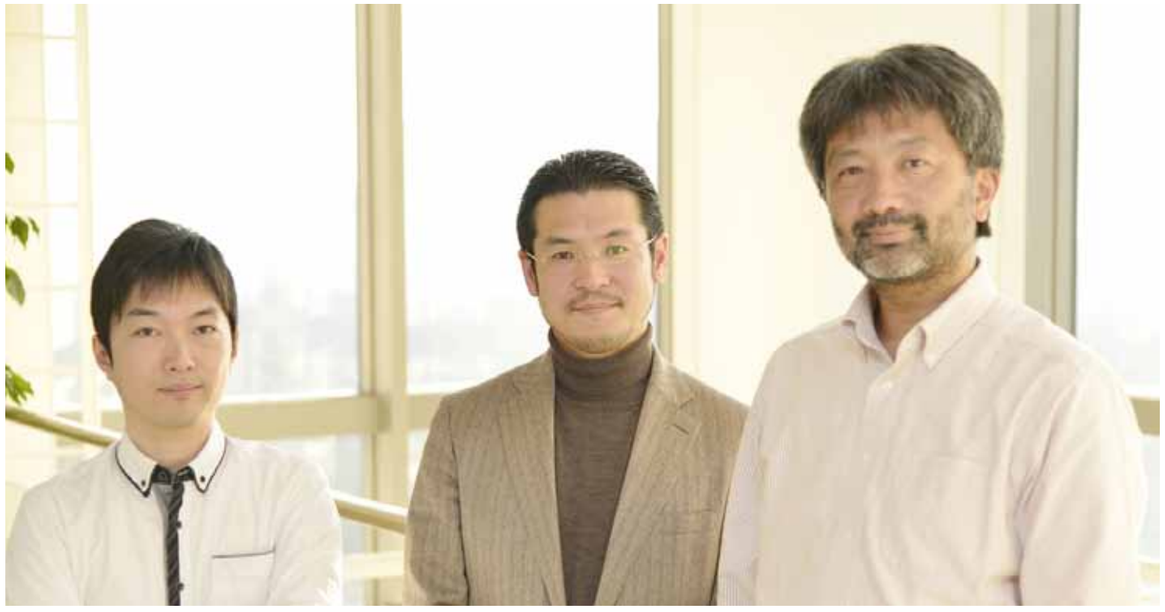
乾　ちなみに、自然言語側も、同じような歴史を辿ってきていますよね。機械翻訳の黎明期には、第二次世界大戦中にコンピュータを使って暗号を解読できたのと同じアナロジーを用いれば、コンピュータを使った翻訳も簡単にできるだろうと楽観視していました。しかし、実際には非常に困難であることがわかったのです。その後、映像とまったく同じように90年代前半に大規模な言語データを活用するとともに、人間が正解を記述したデータを与えて学習させることで、構文解析や浅い意味解析のモデルをつくったのですが、これが大成功したわけです。さらに近年では、WebやSNSの登場によりテキストデータがものすごい勢いで増えており、それらの膨大なデータからさまざまな言語知識や世界知識を取り出していくという方向に向かっています。膨大な量の単語や文章があれば、その組み合わせにより、人間でいう所の「常識」を言語解析に取り込むことができ、ひいてはさまざまな場面における会話の省略などを機械でも補えるようになるかもしれません。

自然言語解析の進展には映像・画像解析も不可欠

乾　そうした中、つねに我々が歯がゆい思いをしていることがあります。自然言語処理ではシンボルの世界の中だけで知識を得ようとする一方で、人工知能（AI）の分野などで議論されているのが、「シンボルをどのようにして実世界におけるその意味と結びつけられるか」という、シンボルグラウンディングの問題です。シンボルだけを取り扱っても、本当の知能を実現できたことにならない可能性がある。人間の知能は多様な外部関係とのインタラクションの中で培われる、と。例えば、人間の子供も母親に「象がいるよ」と教わり、子どもも「あれが象なのだな」と認識します。そうした外部とのインタラクションの中で、さまざまな知能が培われていく。自然言語処理の側にも、シンボルだけを取り扱うことの是非を心のどこかに抱えており、環境とのインタラクションに知能の本質があるのでは、という議論がつねにあります。そうした外部環境とのインタラクションにおける重要な要素の1つが映像・画像であり、それらと言語をペアにして扱うことで、機械にも人間のような追体験をさせることができ、やがてはAIのような進化へとつなげていけるのではないでしょうか。

宮尾　私も、大量のテキストデータから得られる知識と、画像のようにまったく異なるメディアから得られる知識とでは、それぞれが完全に重なるのではなく、実は違うものを捉えている可能性があると感じています。つまり、テキストからだけでは得られない知識があるのではないか、と。いまや、広範な知識や常識を活用して、コンピュータに意味解析させる段階にきており、言語処理だけでなく映像・画像処理も組み合わせることで新しい知見が得られるに違いない。実際に、互いに活用できるリソースもあります。





そのことが、私が映像・画像解析に興味をもったきっかけとなっています。

乾　自然言語側からは、映像・画像解析が次の研究の柱の1つになりうると考えており、まさに宮尾先生がおっしゃったように両者に共通する部分が、これからの研究テーマになっていくのでしょうね。

適切な評価基準をもったタスク設定が不可欠

佐藤　私たちの側からも映像・画像にシンボルを対応づけさせるにあたり、シンボルグラウンディングの問題が挙げられています。映像・画像認識では、映像・画像中の物体や状況などをシンボルに対応付けるところまでが目的であり、シンボルグラウンディングなど意味の問題にはそれほど深入りする必要はないと考えられていました。しかし、学習のためのデータセットを人手で作成する場合の曖昧性、シンボルの語彙数を実用規模にまで拡大してシンボル間の親子関係への対応など、そこかしこに意味の問題が立ちはだかることがわかり、結局、意味の問題に深く立ち入らないとうまくいかない場面が多々現れています。一方で、自然言語処理も、初めにシンボルがあるとはいえ、映像・画像解析側とは逆向きに同じ到達点へと向かっているように感じます。そうした中で、お互いに上手く歩み寄れるような、適切なタスクを設定することで新たに見えてくる知見があると思っています。

乾　例えば、タスク設定として、ラベル付けの問題がありましたね。「これは何の画像なのか」という問いに対して、例えば10種類あるカテゴリ、つまりラベルの中からどれを選べばよいかというものなので、何が正解かは比較的明確です。一方で、キャプションや要約などを生成する問題の場合は、映像・画像、自然言語ともに正しい評価が難しい面もあります。それをどうやって評価するのか。評価の基準がきちんと定められなければ、研究を上手く進めていくことは困難です。また個別技術の追究だけでは成果が出しにくく、研究者のモチベーションも上がりません。

佐藤　おっしゃる通り、映像・画像解析のトップ会議でもその話題が出始めています。研究を次のフェーズへと進めていくためには、適切な評価基準と共通のタスク設定を定めていかなければなりませんね。

宮尾　一方で、例えば「BLEU」という機械翻訳の自動評価尺度があり、最近では、映像・画像解析の研究者の方々も使い始めているようですが、その指標に基づいて正解率を上げる方向へと研究が進んでしまうという懸念もあります。もちろん、そのおかげで機械翻訳の精度も上がっているという側面もありますので、注意深く進めなければなりませんね。

佐藤　同感です。いまやコンピュータで、テキストや映像・画像を含めた外部環境から与えられた情報を認識し、その意味を解析するということが可能になりつつあります。将来的には映像・画像検索や自動キャプション生成をはじめ、監視カメラでのモニタリングや解析など、さまざまな応用例が登場してくると期待しています。そうした中で、先述のように、その過程で生じる個々の問題をブレイクダウンするとともに、研究を確実にステップアップさせていくためのタスク設定を行っていく必要がある。これは映像画像側、自然言語側と両方で取り組んでいかなければならないでしょう。そこには、いずれAI研究も関わってくるのではないでしょうか。楽しみですね。

（取材・文＝伊藤秀樹）

NII essay

写真の歴史性

北本朝展
KITAMOTO Asanobu
国立情報学研究所 コンテンツ科学研究系 准教授

写真とは歴史の記録である。世界のいまを切り取って焼き付けたものが写真なのだから、未来から見ればそれは過去の貴重な記録となる。まあ、当然と言えば当然のことではあるが、最近はこのことの意味を、もう少し深く考えている。

写真は二度と戻らない風景の記録である。そんなことを痛感するきっかけとなったのが、1993年の北海道南西沖地震である。私は大地震の2年前に北海道の奥尻島に旅行し、ウニを拾ったり民宿に泊まったりしながら、楽しい数日を過ごした。ところが、大地震を報じるテレビニュースは、あの町が津波で壊滅して炎に包まれる様子を映していた。あの風景は二度と見られないのか、という衝撃が、風景を写真に残すアーカイブに興味を持つきっかけとなった気がする。

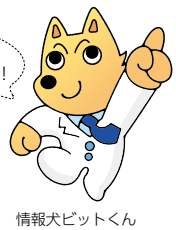
とはいえ、写真の歴史性は、大きなイベントにのみ宿るものではない。日常生活の写真に記録された歴史の断片は、過去の記憶を呼び覚ますきっかけにもなる。ツイッターにはフォロー数200万という超人気アカウント @HistoricalPics があるが、このアカウントがやることは古写真とそのタイトルをつぶやくだけ。人気の秘密は、たまたま投稿された一枚の写真が、すっかり忘れていた記憶を鮮やかに甦らせ、そこから語りが生まれてくるからだろう。

ただ、写真の歴史性は、実はもっと深いのである。そのことに気付いたのは、モバイルアプリ「メモリーハンティング」を開発しているときだった。このアプリは、カメラファインダー上に過去の写真を半透明で重ねることで、過去の写真と同一構図による写真の撮影を支援するものである。こうした同一構図の写真「ビフォーアフター画像」は、防災であれば被災から復興までの変化を記録するメディアとして有用である。

しかし同一構図の撮影は、やってみるとなかなか難しい。撮影者と同一の場所に立つだけでは不十分で、実は姿勢のレベルでも同一性が必要になる。撮影者がしゃがんで撮影した写真と同一構図にするには、自分もしゃがまなければならない。そこで、ふと気づいたのである。同一構図での写真撮影は、実は過去そこにいたはずの撮影者と時間を越えて身体を重ねる体験なのではないかと。もし今は亡き人が撮影した写真があれば、そこに自分の身体を重ねてみて欲しい。それは特別な感慨を呼び起こす体験ではないだろうか。

つまり写真とは、撮影者の身体の痕跡という歴史性も記録したメディアなのである。皆さんが何気なく撮影している写真も、未来の誰かにとっては価値ある写真になるかもしれない。写真の歴史性に思いをはせながら、歴史の中で生きる自分はどんな記録を残せるか、考えてみてはいかがだろうか？

情報から知を紡ぎだす。
NII
国立情報学研究所ニュース [NII Today] 第67号 平成27年3月
発行：大学共同利用機関法人 情報・システム研究機構 国立情報学研究所
〒101-8430 東京都千代田区一ツ橋2丁目1番2号 学術総合センター
発行人：喜連川 優 監修：佐藤一郎
表紙画：城谷俊也 写真撮影：川本聖哉 / 佐藤祐介
デスク：田井中麻都佳 制作：ノーバジェット株式会社
本誌についてのお問い合わせ：総務部企画課 広報チーム
TEL：03-4212-2164 FAX：03-4212-2150 e-mail：kouhou@nii.ac.jp



情報犬ビットくん
(NIIキャラクター)

http://www.nii.ac.jp/about/publication/today/