マルチモーダル空間認識技術

Multimodal Space Recognition Technology

武　井　宏　将

要　約　本稿では，実環境から取得する空間センシングデータを利用した空間認識技術の研究および課題への適用事例について紹介する．空間センシングデータとは，視覚情報を表現する画像情報や三次元空間を表現する三次元点群等を指す．これらの空間センシングデータの活用には，マルチモーダル処理が有効な役割を果たす．マルチモーダルとは，複数のデータソースを利用した処理を指す．ヒトが，視覚，聴覚，触覚，味覚，嗅覚の五感を駆使して様々な判断をしているように，複数の空間センシングデータを駆使して活用することで，単一ソースでは実現できない処理や，単一ソースで処理するよりも精度の高い処理を実現することができる．

2019 年日本ユニシスは，人間の認識・判断を再現する空間認識プラットフォーム BRaVS をリリースした．BRaVS は，マルチモーダル空間認識処理を行うために，ワンパッケージで画像や三次元データ・音を解析できる構成となっている．

今後，マルチモーダル空間認識技術の確立に向けて研究を進めるとともに，BRaVS を通じて技術提供をすることで，様々な課題解決に貢献していく．

Abstract　In this paper, we introduce our research and case study on spatial recognition technology using spatial sensing data acquired from the real environment. Spatial sensing data refers to image information, three-dimensional point clouds, and so on. Multimodal processing plays an effective role in utilizing these spatial sensing data. Multimodal refers to processing using multiple data sources. Human beings have a variety of senses, including the senses of sight, hearing, touch, taste and smell. We make full use of multiple spatial sensing data as if we were making various decisions using our five senses, we enable processing that cannot be done with a single source and provides a higher accuracy than with a single source processing can be realized. We have applied multimodal spatial recognition processing techniques to various problems.

In 2019, Nihon Unisys released BRaVS (Bridging Real and Virtual Space), a spatial recognition platform that reproduces human recognition and judgment. BRaVS provides multimodal spatial recognition processing, including images, 3Ds and sounds.

In the future, we will continue our research to establish multimodal spatial recognition technology, we will provide a new technology through BRaVS. We hope to contribute to solving various problems.

1. は　じ　め　に

ヒトは，実環境の中で視覚，聴覚，触覚，味覚，嗅覚の五感を駆使して日々様々な判断をしながら生活をしている．デバイス技術の進展により，実環境の空間情報をデジタル化して取得できるようになり，情報の精度も年々上がり続けている．取得した実環境の空間情報を空間センシングデータ，それを活用する技術を空間認識処理技術と呼ぶ．

空間認識処理技術において，マルチモーダル処理が有効な役割を果たす．例えば，ヒトが監視を行う場合を例に考える．監視業務では，ヒトが目視で何かおかしなことが発生していないか確認し，異常があれば連絡・注意喚起・救助等を行う．このとき，判断に視覚情報（画像）を利用しているが，同時に聴覚情報（音）や嗅覚情報（におい）も利用している．また，経験によるヒトの記憶を利用している場合もある．このように，ヒトは複数の情報を照らし合わせて判断をしている．コンピュータにより空間認識処理を行う場合も，複数のソースを利用して判断するマルチモーダル処理により，単一ソースでは実現できない処理や，単一ソースで処理するよりも精度の高い処理を実現することができる．

筆者らは，空間認識処理におけるマルチモーダル処理技術の研究と活用を行っている．また，2019 年に人間の認識・判断を再現する空間認識プラットフォーム BRaVS（Bridging Real and Virtual Space）[1]をリリースした．BRaVS は，マルチモーダル空間認識処理を行うために，ワンパッケージで画像や三次元データ・音を解析できる構成となっている．

本稿では，マルチモーダル空間認識処理に関連する技術動向と BRaVS について解説し，マルチモーダル処理を活用した成果を紹介する．2 章では，空間センシングデバイスと深層学習の進展について解説する．3 章では，BRaVS の商品コンセプトを解説する．4 章では，筆者らがこれまで行ってきた，様々な分野にマルチモーダル空間認識処理技術を活用した成果について紹介する．5 章では，今後の展望について述べて，まとめる．

2. センシングデバイスと深層学習の進展

空間認識処理には，実環境の情報をデジタル化する必要がある．空間情報を，どのようなデータとしてどの程度の精度でデジタル化できるかは，実現性の可否や処理の精度に大きく影響する．センサーデバイスの進展は，空間認識処理の進展に非常に重要である．

また，ヒトは日常，ルールに落とすことが難しい，経験に基づく判断を多数行っている．実環境データを扱う空間認識処理においても，そのような処理を行うケースが度々ある．これは，従来コンピュータが不得意としていた処理であるが，深層学習の進展により高精度で実現できるようになってきた．また，深層学習に使われるニューラルネットワークの構造は，マルチモーダルという観点においても非常に有用である．複数のソースをニューラルネットワークに接続できれば，それらを同時に扱う処理が実現できるからである．

本章では，2.1 節にて空間センシングデバイスの進化について，2.2 節にて深層学習の進化について解説する．

2.1. 空間センシングデバイスの進化

本節では，実環境から三次元情報を取得する空間センシングデバイスとして，画像と三次元情報の両方を同時に取得する「RGBD カメラ」と，広い環境の三次元情報を取得する「中長距離 LiDAR」について解説する．

一般的な RGB カメラで撮影した画像が実環境の距離情報を持たないのに対し，RGBD カメラでは，RGB 画像の各ピクセルに深度情報（D）を持つデータを取得できる．RGBD カメラとして，Intel 社が販売している RealSense[2]（図 1 左）や Microsoft 社が販売している Azure Kinect[3]（図 1 右）がある．

図 1　RGBD カメラ （左）Intel RealSense（右）Azure Kinect

深度情報を三次元座標値に変換することにより，RGBD カメラで取得したデータの処理においては，画像処理と三次元データ処理の両方を組み合わせることができる（図 2）．また，2020 年現在，スマートフォンにも RGBD カメラが搭載されはじめている．以前は，実環境から三次元データを取得しようとすると専用のデバイスが必要であり，手軽に利用できるという状況ではなかった．しかし，2000 年に携帯電話にカメラが搭載されてから画像の撮影が非常に手軽になったように，スマートフォンに RGBD カメラが搭載されることで，今後三次元データを手軽に撮影できるようになり，その活用が広がることが期待される．

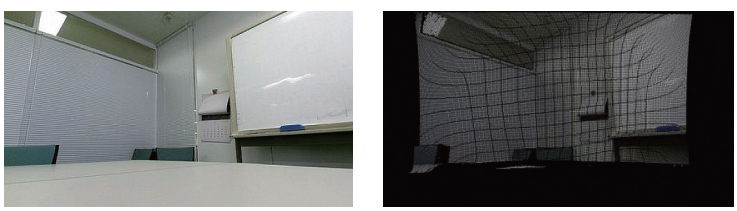

図 2　同時に撮影した二次元画像と三次元点群（左：二次元画像，右：三次元点群）

画像は，テクスチャ（模様）情報が豊富で，あるピクセルの周辺は様々な色の組み合わせを取ることができ，その点の特徴を表現できる．しかし，長さや奥行きといった空間情報は正しく持たない．一方，三次元点群は正確な空間情報を持つ．しかし，実形状の多くの部分は平面に近い形状のため，ある点の周辺は，局所的に平坦に近い形状しか取ることができず，その点の特徴を表現することが困難である．このように画像と三次元点群は互いに双補完的な特徴を持つ．よって，画像と三次元点群の二つのデータを組み合わせることで，お互いに得意な部分を活かした処理ができる．4 章で，三次元 CAD データと画像を利用して画像から空間情報を取得する技術（4.1 節　カメラ画像と 3D CAD データによる三次元空間内動作認識技術）と，RGBD カメラで撮影したデータにおける三次元平面抽出を高速に行う技術（4.2 節　画像と点群を利用した高速三次元平面推定技術）を紹介する．

もう一つの空間センシングデバイスの進展として，中長距離レーザセンサー（以降，中長距離 LiDAR：Light Detection and Ranging）の低価格化について述べる．中長距離 LiDAR は，野外の撮影や工場のデジタル化を目的とした測定用途で利用されてきたが，利用される業務は限られていた．中長距離 LiDAR が注目を集めるきっかけとなったのが自動運転である．自動運転車に取り付けられた LiDAR が，周辺の 3D 情報を取得して，判断をしながら車を走らせている．これにより，一般に LiDAR が知られる存在となった．

LiDAR の低価格化の動きは，半導体型 LiDAR（solid state LiDAR）の登場で進みはじめた．もし，自動運転車が一般に普及した場合，大量生産により，半導体型 LiDAR の価格は数万円程度まで下がると予測されている．

中長距離 LiDAR の測定距離は，一般的に数 100m である．図 3 は，社内の会議室を半導体型 LiDAR により撮影した三次元点群である．このようなデータは 1 台のデバイスで 1 箇所での撮影により取得できる．

図 3　中長距離 LiDAR により撮影した三次元点群

RGBD カメラは，画像と三次元点群を同時に取得できる一方で，測定距離は一般的に 10m 程度である．実課題に適用しようとすると測定距離がネックとなるケースが多々ある．中長距離 LiDAR は，取得できるのは三次元点群のみであるが，広範囲のデータを取得することができるので，利用用途によって使い分けるのが良策である（表 1）．

表 1　RGBD カメラと中長距離 LiDAR の特性比較

	取得データ	測定距離
RGBD カメラ	画像と三次元情報を同時取得	10m 程度
中長距離 LiDAR	三次元点群のみ	100m 程度

2. 2.　深層学習の進化

深層学習が一般に注目を浴びたのは，2012 年のコンピュータによる物体認識の精度を競う国際コンテスト，ImageNet Large Scale Visual Recognition Challenge 2012[4]であろう．このとき，それまであまり利用されてこなかった深層学習による手法が，既存手法から 10%近く正解率を向上して優勝したことで，深層学習への注目が一気に集まった．その後，画像に対する深層学習の研究は進み，画像認識の精度が年々向上していった[5]．現在では，画像認識だけでなく，物体検出・セグメンテーション・異常検知・画像生成とこれまで難しいとされてきた画像処理の課題に対しても盛んに研究が行われ，従来よりも高い精度で実現できることが示されている（図 4）．

深層学習は，ニューラルネットワークと呼ばれる構造を利用したアルゴリズムの総称である．ニューラルネットワークとは，入力層にデータを入力し，隠れ層で順々に計算された結果が，出力層に出力され，その値により判定を行う手法である（図 5）．

ここで，深層学習とマルチモーダルの関係性について述べる．ニューラルネットワークは入力層から出力層までのつながりを表現することができれば計算ができる．そのため，入力を複数にして，それぞれの計算結果を結合するような構造も比較的簡単に構築できる（図 6）．これは，マルチモーダルな入力を一つの処理の仕組みの中で扱うことができることを意味している．例えば単一ソースでは識別が難しいが，複数のソースを利用すれば識別できる問題に効果を発揮する．4 章にて，数値情報と画像を用いた深層学習により橋梁の劣化要因や健全度を推

図 4　深層学習を利用した画像認識処理技術の例

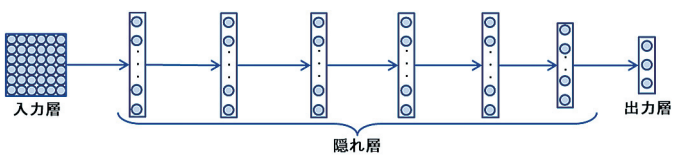

図 5　ニューラルネットワークの概念図

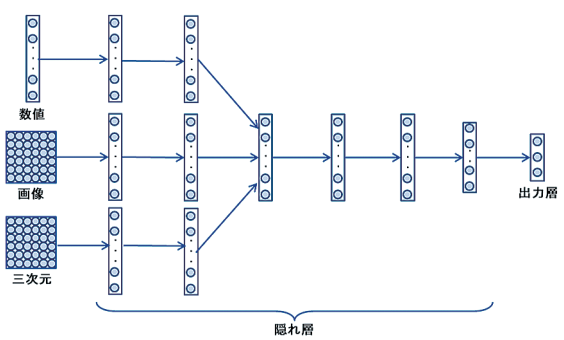

図 6　マルチモーダルなニューラルネットワーク処理

定する技術（4.3 節 画像と諸元情報を組み合わせた橋梁劣化 AI 診断技術）と，画像と数値を用いた深層学習の物理シミュレーションへの適用事例（4.4 節 深層学習＋数値＋画像による CAE 技術）を紹介する．

3.　BRaVS（Bridging Real and Virtual Space）

日本ユニシス株式会社（以降，日本ユニシス）は，画像処理・3D データ処理・深層学習 / 機械学習をこれまで様々な領域に適用し，技術を蓄積してきた．これらの技術を基に，人間の認識・判断を再現する空間認識プラットフォーム BRaVS（Bridging Real and Virtual Space）を 2019 年 5 月にリリースした．BRaVS は，マルチモーダル空間認識処理を行うために，ワンパッケージで画像や三次元データ・音を解析できる構成としている．

空間認識処理を活用するための要素は，大きく「デバイスによる空間センシング」⇒「データ収集」⇒「空間認識処理」に分かれる．BRaVS ではこの流れを早期に実現するための仕組

みとして，深層学習・画像処理・3D データ処理・音響処理をパッケージングした「BRaVS Library」と WebAPI 公開基盤である「BRaVS Platform」を提供している（図 7）．

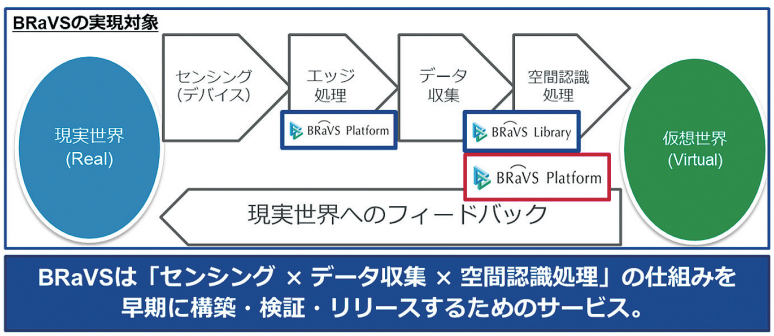

図 7　空間認識処理活用のフロー

マルチモーダル処理には，ヒトの視覚・聴覚・空間把握に対応する各ソースを一元的に扱えるパッケージが要求される．BRaVS Library は，マルチモーダル空間認識に用いる処理を一元的に扱うことができるライブラリである（図 8）．

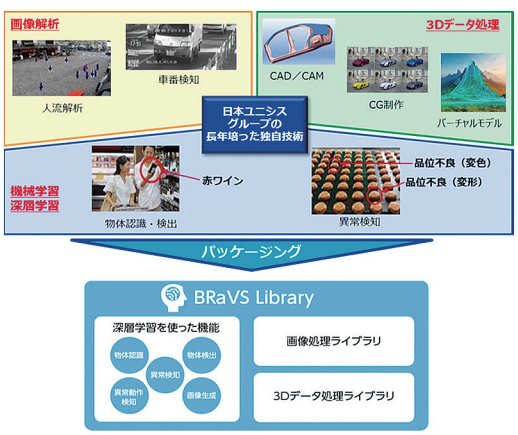

図 8　BRaVS Library

日本ユニシスは，BRaVS を通じてマルチモーダル空間認識処理技術を提供していく．また，筆者らの研究により作られた空間認識処理技術を BRaVS に組み込むことで，多くの方々の課題解決のために，研究成果を適用していきたい．

4.　事例紹介

本章では，マルチモーダル空間認識処理技術を活用した四つの事例を，4.1 から 4.4 の各節で紹介する．一つ目は，カメラ画像と 3D CAD データを組み合わせることで，カメラ画像から三次元空間上の動きを把握する技術[6]である．二つ目は，RGBD カメラからのデータを用いた，画像と 3D データを利用した高速平面推定の技術である．三つ目は，橋梁の劣化要因の診断を画像と数値データを組み合わせた深層学習により実現した技術である．四つ目は，物理シミュレーションを画像と数値の深層学習により実現した技術である．

各事例とも，単独のソースのデータのみでは実現が困難または課題が残るのに対して，マルチソースのデータにより課題解決をしたものである．

4. 1.　カメラ画像と 3D CAD データによる三次元空間内動作認識技術

単眼の RGB カメラで撮影した画像は，距離の情報を持たない．本研究では，画像に撮影されている対象と同一の形状の 3D CAD データを利用し，画像内でその形状をトラッキングすることで，その対象の三次元空間上での動きを再現する技術を開発した（図 9）．

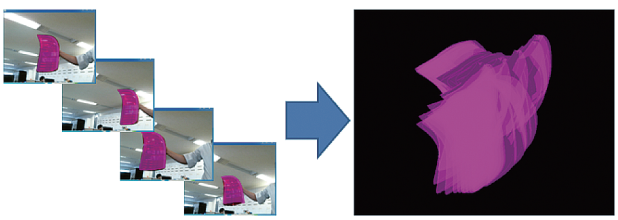

図 9　画像内の対象をトラッキングすることで三次元空間上での動きを再現

単一のビデオ映像内の対象物体をトラッキングする手法は，画像処理の分野で多数提案されている[7]がトラッキング手法の多くは二次元画像内における対象物体のトラッキングであり，本システムで行っている三次元トラッキングは実現できない．単眼カメラを用いて三次元空間を認識する研究として PTAM（Parallel Tracking and Mapping）[8]がよく知られているが，PTAM は空間把握を目的として設計されており，物体追跡に利用するためには三次元空間の認識精度や追跡精度の面で十分でない．本技術は，画像と三次元データをマルチモーダルに利用することでこれらの課題を解決した．

画像内の対象に対して 3D CAD データを当てはめると，画像のカメラ座標系におけるその対象の位置を算出することができる（図 10）．対象をトラッキングすることで，三次元空間上の動きを再現することができる．

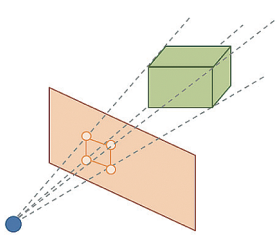

図 10　対象の当てはめとカメラ座標系における位置

本技術は，距離情報を持たない画像情報と大きさの情報を持つ三次元形状を組み合わせることで，動的に三次元情報を取得するものである．

4. 2.　画像と点群を利用した高速三次元平面推定技術

三次元点群から平面を抽出する処理は度々行われる．点群の平面計算について，以下の二つの課題が知られている．

課題 1：隣接頂点の検索にかかる計算量が大きい．

課題 2：近傍点群からの平面算出が不安定.

ここで，RGBD 画像から平面を抽出することを考える．このとき，深度情報（D）を三次元点群に変換し，三次元点群の平面を抽出するというアプローチも考えられる．しかし，その場合上記の課題に突き当たる．そこで，画像（RGB）を活用することで，課題を解決する技術を開発した.

本手法に類似した手法として，Erdogan らの論文[9]で提案された手法がある．我々の手法は，SuperPixel 算出に平面算出のしやすさを考慮したアルゴリズムを利用した点と，領域をマージする処理に繰り返し計算を伴う確率的アルゴリズムを用いず，領域拡張法[*1] により高速なマージ処理を行った点で彼らの手法と異なる.

課題解決のアイデアは，画像を SuperPixel と呼ばれる単位に分割することである．SuperPixel とは近くにある類似した画像を一まとまりにした単位である．SuperPixel を取得する手法として SLIC[10]を用いた（図 11）.

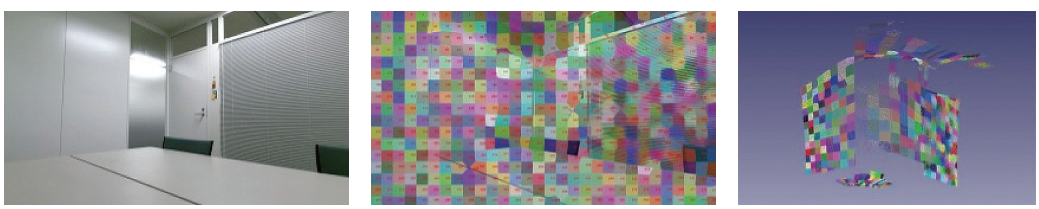

図 11　SuperPixel 抽出と SuperPixel を対応付けた三次元点群

SuperPixel 毎に法線方向を算出し，SuperPixel の近傍情報により周辺の法線との関係を調べることで課題 1 を解決する．三次元点群の頂点から近傍を探していくよりも高速に計算できる．また，SuperPixel は局所的に平面となるため，推定する平面と大きく差異が発生することもない.

次に，SLIC により算出した SuperPixel は，大きさが均等であり，SuperPixel の境界は三次元点群の境界に比較的一致している（図 12）．そのため，SuperPixel が自然と境界条件による頂点の選択の役割を果たすため，安定した精度のよい平面を算出することができ，課題 2 を解決する.

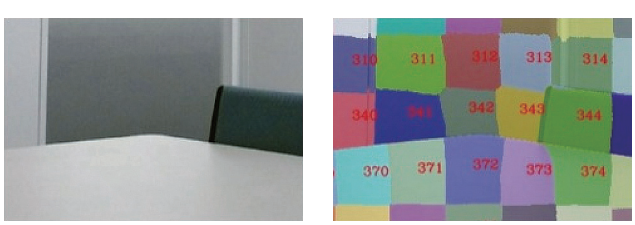

図 12　SLIC を利用した SuperPixel 抽出

本手法は，三次元点群のみ処理をすると発生する課題を，画像（RGB）を補完的に利用して解決するものである.

4. 3.　画像と諸元情報を組み合わせた橋梁劣化 AI 診断技術

従来，橋梁点検ではコンクリート診断士と呼ばれる専門技術者が橋梁を目視で確認して劣化状況を診断していた．橋梁の劣化診断には，劣化の原因を示す「劣化要因」の判定と，劣化の程度を示す「健全性」の判定の二つがある．日本ユニシスは，建設総合コンサルタントの株式会社日本海コンサルタント（以降，日本海コンサルタント）と共同で，深層学習を用いた橋梁劣化診断技術の検証を行った（図 13）．

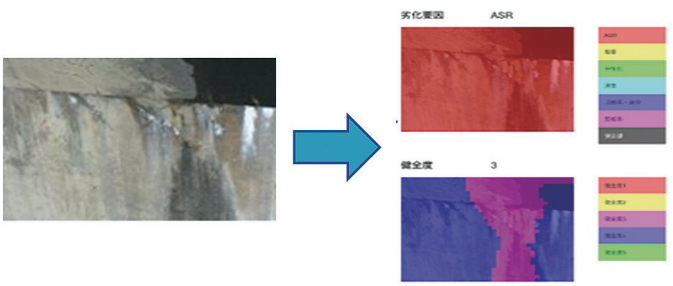

図 13　深層学習を用いた橋梁劣化診断

橋梁診断についてヒアリングした結果，コンクリート診断士は目で見た情報以外にも，地域特性等を加味して判断していることがわかった．そこで，深層学習の入力として，診断士が判断に利用している諸元情報も利用することで精度向上できることを確認した．劣化要因の診断において，諸元情報なしの場合の正解率 73.1％に対し，諸元情報ありの場合，正解率 91.2％となった．また，健全度の診断において，諸元情報なしの場合の正解率 60.1％に対し，諸元情報ありの場合，正解率 83.4％となり，それぞれ大幅な精度向上を示した[11]．この技術は，画像と数値情報をマルチモーダルに利用することで，画像のみでは難しい橋梁劣化診断の精度向上を実現したものである．

インフラ構造物の劣化診断に深層学習を適用する試みは様々行われている．しかし，その多くは画像のみを入力として，傷の抽出のような要因の一種を取得するものであり，橋梁劣化診断のスコアを直接算出する取り組みはほとんどない．本技術は診断士の知見を含めることで，診断士の判断と同等レベルに達した点で他の研究と異なる．

また本技術をベースとして，AI 橋梁診断支援システム「Dr.Bridge」を日本海コンサルタントと共同で，2020 年 6 月にリリースした[12]（図 14）．

図 14　Dr.Bridge

4. 4.　深層学習＋数値＋画像による CAE 技術

CAE（Computer Aided Engineering）とは，コンピュータ上で行う物理シミュレーションのことである．製造業における製品設計で一般的に利用されている．CAE ソルバー[*2]は，高精度化を目指して多くの技術進展がなされてきた．その一方で，CAE は解析にかかる計算量が膨大なため，解析に時間がかかることや，高スペックなマシンを要することが課題として挙げられる．特に設計業務の序盤では，CAE の結果の精度は多少犠牲にしてでも多くのケースを回したいというニーズがあるが，このようなニーズに現状の CAE では応えることが難しい．

筆者らは，CAE ソルバーの進展の流れとは別な方向性として，CAE の各種パラメータや形状の入力と解析結果を深層学習することで，大きく精度を落とすことなく，従来の CAE よりも低スペックのマシンで，高速に結果を取得できる技術を開発した（図 15）．

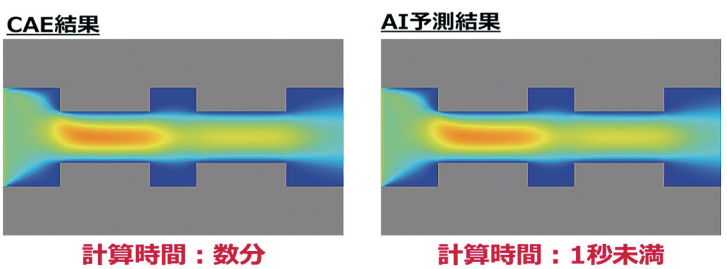

図 15　CAE の結果と AI 予測の結果の比較

深層学習ベースの CAE は従来の CAE と比較して計算が高速であり，マシンもワークステーションと GPU1 台程度で実現できる．製造業における成形解析や流体解析に本技術を用いることについて顧客と検証している．本技術は，画像と数値情報をマルチモーダルで利用することで，実現したものである．

5.　お　わ　り　に

筆者らは，本論文で紹介した事例をはじめとした様々な課題に対して，空間センシングデータを活用したマルチモーダル空間認識処理技術の研究および実課題への適用を実施してきた．マルチモーダル空間認識処理技術についての，今後の展望を述べる．

一つは音の活用である．音を利用したマルチモーダルの可能性については，様々な場面が想定されており，技術研究も進めている．今後，実課題に対して適用することでその有用性を確認していきたいと考えている．

もう一つは，中長距離 LiDAR と RGB カメラを組み合わせたマルチモーダル処理である．中長距離 LiDAR と RGB カメラを組み合わせることで，長距離範囲を測定した三次元点群と対応する画像を取得することが可能となり，より広範囲のデータを利用したマルチモーダル処理を実現できることが期待される．

空間センシングデバイスは年々進化しており，この動向をフォローしつつ，引き続き新たな空間認識処理技術を研究していく．また，研究した技術は，BRaVS を通じて実課題への適用を行い，様々な課題解決に貢献していきたいと考えている．

＊1　ある位置を開始位置として，近傍をたどりながら処理範囲を広げていく手法．Region Growing とも呼ばれる．
＊2　数値計算を実行して物理シミュレーションの方程式を解くプログラムのこと．提供ベンダーとして Ansys 社[13]や Autoform 社[14]が有名である．

参考文献

［1］「BRaVS Library®/BRaVS Platform®」 https://www.unisys.co.jp/solution/tec/iot/bp/bravs.html
［2］「インテル RealSense テクノロジー」 https://www.intel.co.jp/content/www/jp/ja/architecture-and-technology/realsense-overview.html
［3］「Azure Kinect DK」 https://azure.microsoft.com/ja-jp/services/kinect-dk/
［4］「ImageNet Large Scale Visual Recognition Challenge 2012（ILSVRC2012）」 http://image-net.org/challenges/LSVRC/2012/index
［5］「ImageNet: Where are we going? And where have we been?"」 http://image-net.org/challenges/talks_2017/imagenet_ilsvrc2017_v1.0.pdf
［6］ 武井宏将，"ビデオ映像内対象物体の三次元トラッキングシステム"，ユニシス技報，日本ユニシス，Vol.34 No.3 通巻 122 号，2014 年 12 月
［7］ M.Isard A.Blake, "CONDENSATION — conditional density propagation for visual tracking", *Int. J. Computer Vision, Vol.29-1, pp.5-28, 1998.*
［8］ G.Klein D.Murray, "Parallel Tracking and Mapping for Small AR Workspaces", *In Proc. of ISMAR'07, pp.225-234, 2007.*
［9］ C.Erdogan M.Paluri F.Dellaert, "Planar Segmentation of RGBD Images using Fast Linear Fitting and Markov Chain Monte Carlo", Proc. of Computer and Robot Vision, pp.32-39, 2012.
［10］ R.Achanta A.Shaji K.Smith A.Lucchi P.Fua S.Susstrunk, "SLIC Superpixels Compared to State-of-the-art Superpixel Methods", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.34-11, pp.2274-2282, 2012.
［11］ 町口敦志，喜多敏春，多田徳夫，武井宏将，近田康夫，"ディープラーニングによる橋梁の劣化要因・健全度判別モデルの構築"，土木学会 第 74 回年次学術講演，2019 年
［12］ 橋梁点検業務の省力化と品質向上を AI で実現！～ AI 橋梁診断支援システム「Dr. Bridge」提供開始～，プレスリリース，(株)日本海コンサルタント，2020 年 5 月，「橋梁点検業務支援サービス Dr.Bridge」 http://www.dr-bridge.ai/
［13］ アンシス・ジャパン株式会社　https://www.ansys.com/ja-jp
［14］ オートフォームジャパン株式会社　https://www.autoform.com/jp/

※　上記参考文献に含まれる URL のリンク先は，2020 年 7 月 20 日時点での存在を確認．

執筆者紹介　武　井　宏　将（Hiromasa Takei）
2004 年 日本ユニシス(株)入社．入社時より CAD/CAM 分野のシステム開発業務に従事．2013 年より画像処理・三次元形状処理・人工知能の研究開発・技術検証業務・商品開発業務に従事．エバンジェリスト（AI）．