UC Berkeley

UC Berkeley Previously Published Works

Title

IDEAL: Images Across Domains, Experiments, Algorithms and Learning

Permalink

https://escholarship.org/uc/item/0wd1v9h5

Journal

JOM, 68(11)

ISSN

1047-4838

Authors

Ushizima, Daniela M
Bale, Hrishikesh A
Bethel, E Wes
et al.

Publication Date

2016-11-01

DOI

10.1007/s11837-016-2098-4

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

JOM, Vol. 68, No. 11, 2016
DOI: 10.1007/s11837-016-2098-4
© 2016 The Minerals, Metals & Materials Society (outside the U.S.)

IDEAL: Images Across Domains, Experiments, Algorithms and Learning

DANIELA M. USHIZIMA,[1,7,8,9] HRISHIKESH A. BALE,[2]
E. WES BETHEL,[1] PETER ERCIUS,[3] BRETT A. HELMS,[3]
HARINARAYAN KRISHNAN,[1] LEA T. GRINBERG,[4]
MACIEJ HARANCZYK,[1] ALASTAIR A. MACDOWELL,[5]
KATARZYNA ODZIOMEK,[6] DILWORTH Y. PARKINSON,[5]
TALITA PERCIANO,[1] ROBERT O. RITCHIE,[2,7] and CHAO YANG[1]

1.—Computational Research Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA. 2.—Materials Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA. 3.—Molecular Foundry, Lawrence Berkeley National Laboratory, Berkeley, CA, USA. 4.—Memory and Aging Center, University of California San Francisco, San Francisco, CA, USA. 5.—Advanced Light Source Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA. 6.—Laboratory of Environmental Chemometrics, Faculty of Chemistry, University of Gdansk, Gdańsk, Poland. 7.—University of California Berkeley, Berkeley, CA, USA. 8.—e-mail: dushizima@lbl.gov. 9.—e-mail: dani.lbnl@berkeley.edu

Research across science domains is increasingly reliant on image-centric data. Software tools are in high demand to uncover relevant, but hidden, information in digital images, such as those coming from faster next generation high-throughput imaging platforms. The challenge is to analyze the data torrent generated by the advanced instruments efficiently, and provide insights such as measurements for decision-making. In this paper, we overview work performed by an interdisciplinary team of computational and materials scientists, aimed at designing software applications and coordinating research efforts connecting (1) emerging algorithms for dealing with large and complex datasets; (2) data analysis methods with emphasis in pattern recognition and machine learning; and (3) advances in evolving computer architectures. Engineering tools around these efforts accelerate the analyses of image-based recordings, improve reusability and reproducibility, scale scientific procedures by reducing time between experiments, increase efficiency, and open opportunities for more users of the imaging facilities. This paper describes our algorithms and software tools, showing results across image scales, demonstrating how our framework plays a role in improving image understanding for quality control of existent materials and discovery of new compounds.

INTRODUCTION

Emerging technologies in complementary metal-oxide-semiconductor (CMOS) and charge-coupled device (CCD) image sensors have made digital images ubiquitous. Collecting experimental outcomes by keeping images as scientific records has become common practice, but automated algorithms to extract information from such datasets have not evolved at the same pace. The increasing rate at which images are produced, combined with the broad variety of sensors, pose algorithmic challenges when applied to large volumes of data, particularly when images contain heterogeneous structures in multiple scales.

As an example, a material scientist analyzes samples at different angles, granularities, and mechanical conditions to assess microstructures. With the ability to acquire micrographs at high spatial and temporal resolutions, understanding experimental images requires more efficient analysis schemes than manual inspection can provide. A current workaround is to simply downsample datasets, so that existing tools can help curate data. This approach is likely to miss the precision and subtlety provided by modern high-resolution instruments. To address image analysis at scale, we have been constructing coherent, cross-domain computing approaches that take five key components into

(Published online September 12, 2016)

consideration: detectors, images, algorithms, data representation and computing architectures. We organize the analysis of experimental data into three main strategies: one-off, generic, and motif-centric.

First, the most common strategy is to develop a one-off pipeline, in which semi-automatic analysis codes are tailored to deal with a small dataset for a narrow science problem. Quickly deployable and human-dependent, one-off workflows often include tailoring thresholds and hand-tuning filters to individual images before structure measurements. Alternatively, a second path is to adopt more sophisticated, yet generic and fully automated frameworks that enable, for example, calculation of subspace partitions[1] from large image sets.[2] These frameworks often demand longer time to implement, but require less manual interaction. Limitations that may arise include: (1) algorithms perform well on simulated[3] but not real data;[4] and (2) the segmentation inaccuracies undermine measurements at subsequent processing tasks.[5]

Our current approach follows a third path, combining relevant tasks from the previous strategies to account for the high data throughput regime (e.g., terabytes per experiment) while imposing controlled amounts of human labor. By allowing and tracking human interaction while working with one-off pipelines and generic tools, we can record visual cues and mine data collected from the user preferences. These non-image data sources feed both models for simulated data[6] and provide information to create repeating patterns, or motifs, for scientific image analysis.[7]

This paper introduces IDEAL, our framework and project to address Images across Domains, Experiments, Algorithms and Learning. It describes the development of our main methods and their application to materials science, involving microCT of ceramic composites ("Ceramic Matrix Composites" section), scanning electron microscopy (SEM) of nanoparticles ("Toxicity of Nanoparticles" section) and STEM of polymeric films ("Films for Next Generation Microelectronics" section), as summarized in Table I. "High Throughput Microscopy" section outlines common problems such as data representation in multiple domains, and "Pattern Recognition" section describes pattern recognition tools that can be used among different image analysis steps. "Evolving Architectures" section lists some strategies to tackle massive image sets, taking advantage of upcoming hardware technologies. The sections that follow discuss how the synergy of projects in different scales has the potential to bring further knowledge to experimental sciences.

BACKGROUND

In material science, the investigation of a specimen involves examining and making sense of a variety of microstructures and geometric constructions, as well as understanding how certain matter organizations can lead to high-performance configurations. The recent ability to quickly collect large image sets has created new challenges in several areas that rely upon image-centric data. A vision of some scientists is to be able to detect "faces of scientific data" with the same ease as occurs in face recognition.

One obvious area in which pattern recognition/computer vision algorithms have made great strides has been in face identification. Solutions often use a machine-learning algorithm called deep learning, which employs a neural network approach with several processing layers and access to large

Table I. Microscopic images of materials across scales: specifications and methods

Materials	Resolution (μm)	Image modality	Imaging contrast mechanism	Data analysis for specimen quantification
Ceramic composites	0.65–1.3	MicroCT	X-ray attenuation contrast	Detection of fibers, fiber breaks and cracks using graph-based and template matching ML algorithms. "Ceramic Matrix Composites" section. Figure 1
Geological samples	0.65–2.5	MicroCT	X-ray attenuation contrast	Segregation of components from multi-phase specimens using Markov Random Field ML algorithms. "Ceramic Matrix Composites" section. Figure 5
Nanoparticle clusters	0.2457	SEM	Electron scattering	Counting, topographical characterization, morphology, particle distribution, ensemble representativeness. "Toxicity of Nanoparticles" section. Figure 3
Thin films	0.00164	STEM tomography	Electron transmission	Pore organization across film, associated to level of pore coalescence; surface density analysis correlated to dielectric constant measurements. "Films for Next Generation Microelectronics" section. Figure 4

image sets, tagged and untagged. According to Krizhevsky et al.,[8] datasets of labeled images on the order of tens of thousands of samples are relatively small and allow for simple recognition tasks. In this context, ImageNet has become a standard database for benchmarking large-scale object recognition[9] because it offers millions of cleanly sorted and annotated images to train classifiers.

There are several reasons why the classification of scientific images is not yet as evolved as face recognition for natural images. First, the lack of tagged and annotated image data sets is an obstacle for many computer vision and machine learning (ML) methods. Second, while deep learning has improved certain kinds of image classification, including ascertaining "What object is in the scene?",[10] the general tasks of material sciences go beyond identification to include semantics, object relationships, and decision-making based on the situation and priors. Third, there can be a hindrance to data sharing by some projects. And fourth, there can be surprising variability between collected images, even when considering the same instrument.

As an example of the challenges that come from scientific images, the Lawrence Berkeley National Laboratory (LBNL) Advanced Light Source (ALS) microtomography instrument (Beamline 8.3.2) scans a variety of samples, including natural biomaterials such as bone and nacre, and advanced hierarchical structural and functional materials, such as SiC composites and batteries, and several geological samples and life science specimens, such as plants and insects. The number of scans ranges between single shot to time-lapse radiographies, with varying experimental settings (e.g., infiltration, stress) and under different investigations. In collaboration with imaging facilities, such as ALS and the LBNL Center for Advanced Mathematics for Energy Research Applications (CAMERA), we have explored use-cases and developed software tools that cover a set of pattern recognition and analysis problems. A great deal of this work forms the basis for the research and development project called IDEAL, which we introduce in this paper.

There has been interest throughout the community in creating tools that can tackle scientific images. Other efforts besides ours include (1) tools at the ANL Advanced Photon Source, including Tomopy[11] for image reconstruction[12] and Midas[13] for analysis of grain interrelationships in crystalline materials; (2) work on the Materials Knowledge System (MKS), led by researchers at Georgia Tech, supports multi-scale materials science investigations using python packages that enable a range of functions, from synthetic data construction to spatial statistics; and (3) PyHST2, from the European Synchrotron Facility (ESRF), which exploits hybrid architectures using both central processing units (CPUs) and GPUs to deliver parallel processing techniques. While our multidisciplinary teams leverage some of these tools for the data reconstruction, IDEAL apps are focused on recognizing geometrical structures and measuring material deformation from 3D structured meshes, like image stacks, as the initial point. The next section discusses some use-cases that apply our IDEAL software stack.

ACROSS LENGTH SCALES

By selecting specific image problems to balance depth and breadth in engineering science domain applications, we have deployed key methods that work in multiple scales. This section describes three cases: (1) analysis of advanced hierarchical woven ceramic fiber matrix composites for hypersonic flight applications, which are being developed through iterative improvements of material properties; (2) standardization of experimental records using morphometric algorithms to analyze nanoparticle microscopy, which correlate toxicity to nanoparticle morphology; and (3) quality control of polymeric films with application to microelectronics to minimize the dielectric constant for mesoporous organic composites.

Ceramic Matrix Composites

Ceramic matrix composites (CMCs) provide exceptional strength-to-weight ratio capabilities, appropriate to the construction of the next generation of jet engines. For such a purpose, exposure to high temperature and microstructural mechanisms may require metal replacement by reinforced CMC.[7,14,15] SiC-SiC composites are a type of CMC, and the samples described in this paper were fabricated at Hypertherm (Huntington Beach, CA, YSA) by weaving bundles of SiC fibers (Fig. 1), followed by a chemical vapor-infiltrated SiC matrix.[16] The goal is to monitor material resistance for high-performance design, with a broad array of applications at LBL, UCB, Air Force, Teledyne, General Electric, and NASA.

In situ mechanical testing to evaluate the material behavior often involves subjecting samples to incremental tensile load under a simulated high-temperature environment during 3D data acquisition. Using hard x-ray micro-tomography (microCT), the resulting images represent a 3D map of the x-ray attenuation coefficient of the materials within the sample. Inspection involves three key factors: (1) identification of the different components of the sample using graph-based ML schemes;[17] (2) detection of microcracks associated to specific conditions of strain and temperature; and (3) location of fibers, and pull-outs within the material.[16]

In order to speed up image analysis of CMC microCT data, we have designed tools that provide (near) real-time feedback of post processed or in situ processed images. As an example, we have implemented F3D,[7] a platform-portable library that ensures these algorithms are usable by a wide

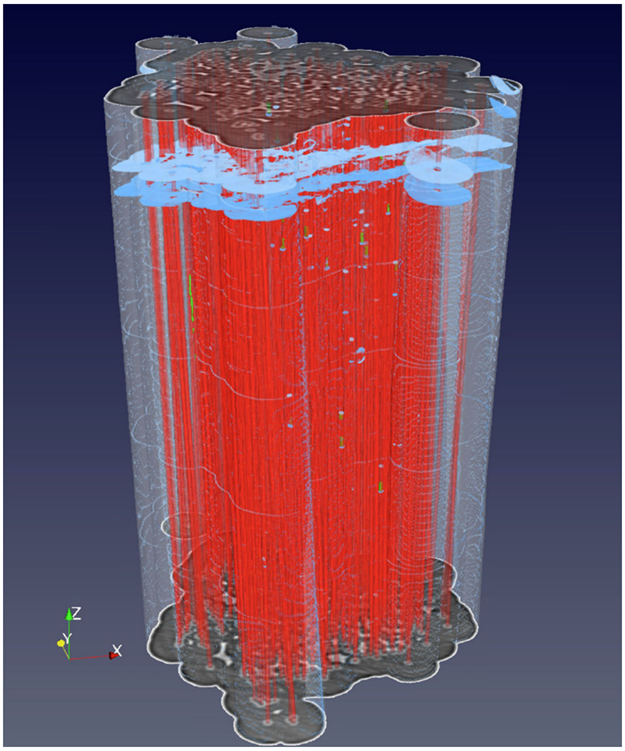


Fig. 1. Volume rendering using paraview of a SiC ceramic composite after segmentation process showing detected SiC fibers (red), detected matrix cracks (blue) and detected fiber breaks (green).

range of users. F3D contains image processing kernels, written in OpenCL, that accelerate analysis of GB scale data through GPU-aware code. Previous work[7] showed 30 GB image stacks being processed in 2.4 min, using a computing system with a Intel(R) Xeon(R) CPU E5-2660 at 2.20 GHZ, 62 GB RAM, and 3 NVIDIA Tesla K20X GPU. Figure 2 shows the user interface of our tool, used in both the analysis of CMC and the characterization of geological samples, whose specifications are shown in Table I.

F3D* is a 3D plugin to ImageJ/Fiji,[18] and it hides parallelism complexity while providing algorithms to enhance particular patterns that work as landmarks during fiber detection, and deformation, such as breaks and ceramic matrix cracks; these are crucial for microstructure characterization. Our algorithms rely, in part, on linear and non-linear transformations using 3D structuring elements, e.g., cylinders.

As an example, we use fiber two-dimensional (2D)-profiles (ellipses) as input to a template matching (TM) algorithm[19] to detect tridimensional fibers, although TM is a parallel-stacked 2D-centered code. Initially, the user selects a few examples or prototypes that represent the expected fiber profiles. Next, our algorithm runs two main calculations several times to find the mean square error (MSE) and the normalized cross-correlation coefficient (NCC). The MSE determines the most suitable prototype p and the NCC computes the similarity between each pixel of the image and p. The "learning" associated to this approach lies on the two recognition tasks: (1) defining superpixels through statistical region merging,[4] and (2) finding the most likely pattern (fiber profile) through TM, given a set of prototypes.[20]

The MSE and CC are defined as

$$\mathrm{MSE}(x,y) = \frac{1}{n}\sum_{i,j}(p(i,j) - f(x+i,y+j))^2 \qquad (1)$$

and

$$\mathrm{CC}(x,y) = \frac{\sum_{i,j}(p_{ij} - \bar{p})\sum_{i,j}(f_{ij} - \bar{f})}{\left[\sum_{i,j}(p_{ij} - \bar{p})^2\sum_{i,j}(f_{ij} - \bar{f})^2\right]^{\frac{1}{2}}} \qquad (2)$$

where i, j are image indexes, $f_{ij} = f(x+i, y+j)$ for simplicity, $\bar{\ }$ is the mean value of $(\cdot)$, and p is the pattern prototype to be found in f. The values of MSE range between [0, 1] and the values of CC between $[-1, 1]$.

Toxicity of Nanoparticles

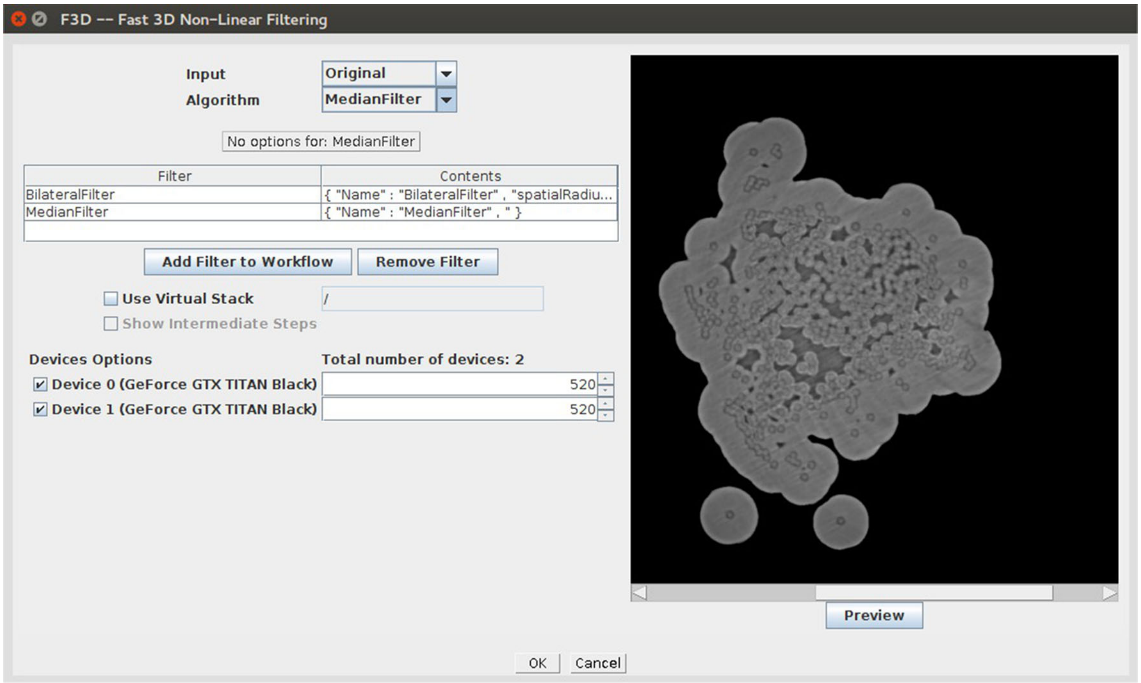
While the toxicity of bulk materials is affected mainly by their composition, for nanoparticles (NPs), the properties of clusters such as size, surface area, morphology (shape), and surface topography are preponderant factors. Depending on such properties, NPs engage in different types of interactions with living cells or tissues, and consequently various toxicity mechanisms are at play.

Figure 3 shows tricalcium phosphate (TCP) $\mathrm{Ca_3(PO_4)_2}$, a naturally occurring mineral with a wide range of biomedical applications. Using a Phenom ProX Desktop SEM (accelerating voltage: 5000, 15,000 V), we obtain grayscale (8-bit) images of TCP grains. At ×400 magnification, the 2048 by 2048 pixels micrographs are scaled to 3.061 pixels/μm (0.3267 pixels/μm).

One of the challenges of working with NPs is the fact that they act like neither bulk molecules nor single molecules, and instead behave like something in between, and hence conventional modeling tools have limited utility. By extracting NPs features from the SEM images, we can incorporate them into a quantitative structure–activity relationship (QSAR) models in order to predict NP toxicity[21] or physico-chemical properties.[22] QSAR methods employ combinations of descriptors in the prediction of physico-chemical and/or biological properties of chemical substances, based on statistical models derived from measured data.[23] QSAR modeling for nanoparticles ("Nano-QSAR") focuses on estimating NP biological effects.

To that purpose, we have developed special tools for analyzing SEM, using unsupervised classification for NP segmentation and feature extraction procedures for automated quantitation.

*https://github.com/CameraIA/F3D.



F3D -- Fast 3D Non-Linear Filtering

Input: Original

Algorithm: MedianFilter

No options for: MedianFilter

Filter	Contents
BilateralFilter	{ "Name" : "BilateralFilter" , "spatialRadiu...
MedianFilter	{ "Name" : "MedianFilter" , " }

☐ Use Virtual Stack /

☐ Show Intermediate Steps

Devices Options — Total number of devices: 2

☑ Device 0 (GeForce GTX TITAN Black) 520

☑ Device 1 (GeForce GTX TITAN Black) 520

Fig. 2. Graphical user interface of the F3D (fast 3D non-linear filtering) plugin: available parameters on the left and preview image on the right. Parameters can be easily tested on small subsets of the original image allowing quick evaluation of settings before batch processing.

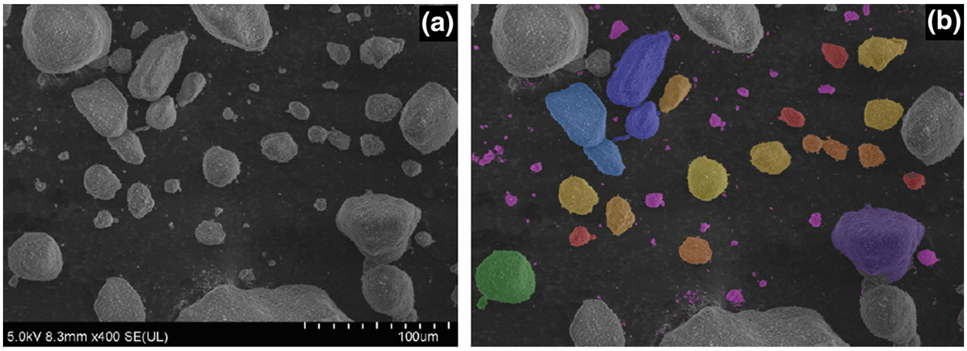


Fig. 3. SEM of tricalcium phosphate $Ca_3(PO_4)_2$ agglomerates during NP analysis to determine toxicity: (a) original SEM and (b) color-coded NPs by surface area.

Additionally, we have built algorithms for estimating the minimal number of images required to retain the NPs ensemble properties, based on a specific feature set; our code selects a smaller subset of the images to be kept on record, which represents the ensemble statistically.[24] The "learning" associated to this approach lies on using clustering (e.g., k-means) and classification (e.g., SVM), for the recognition of the most representative subset of SEM images given NPs descriptors.

Films for Next Generation Microelectronics

Controlling the material porosity can be useful in revealing and fine-tuning its properties as a dielectric, sorbent, or active layer for applications in catalysis, health, and energy. Pores with mesoscale dimensions are of particular interest in the design of periodic mesoporous organosilicas (PMO) thin films. By embedding molecular or polymeric porogens within the host material, mesopores can be controlled during the material creation. Mesopore dimensions follow specific design rules regarding shape, spatial arrangement, and defect structure, which together enable the assembly of well-controlled, ordered architectures. This section describes some of the tools we have built to assess the factors governing porogen packing and shape persistence during mesoscale assembly.

In,[25] we formulated image attributes that can be used to understand the fundamental packing limits for spherical block copolymer (BCP) micellar porogens during the assembly and thermal processing of PMOs. Images consist of scanning transmission electron microscopy (STEM) tomography of material samples, presenting either ordered or disordered

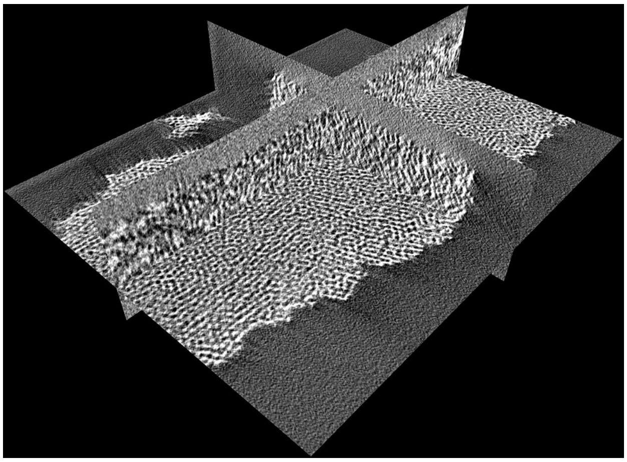


Fig. 4. STEM-tomography obtained on a grain of mesoporous organosilica (58% porous) flaked off from a Si substrate after thermal processing. The films were prepared from 11-nm PS-b-PDMA micelles.

domains in 3D space, as shown in Fig. 4. We have defined and deployed a set of statistical descriptors for STEM images that indicates pore packing relationships and pore organization throughout the film. As a result, these indicators correlate pore coalescence to dielectric constant measurements.

The coalescence indicators for STEM of PMO films explore pore packing by calculating gray-level variations using texture analysis.[26] The pore architecture information is obtained by local variations in image intensity, which is too fine to be distinguished as separate objects by the observer. The core algorithm involves the joint probability distribution of a gray-level image I_g at every two pixels, $i = I_g(x,y)$ and $j = I_g(\hat{x},\hat{y})$, conditioned on θ and d, the direction and distance, respectively. In other words, we calculate the Gray-Level Co-occurrence Matrix (GLCM) to measure the spatial organization of the pixel intensities, using the following equations: $q(i,j|d,\theta) = \#(i,j) \in I_g$, such as $j = \rho(i|d,\theta)$, for a pixel j at position $\hat{x} = x + d_1, \hat{y} = y + d_2$, where d_1 and d_2 take values among $\{-d, 0, d\}$ depending on the direction $\theta \in \{0, 45, 90, 135\}$. We write GLCM as $q(i,j|d,\theta)$ or $q_{i,j}$, for simplicity, and the symbol $\#$ indicates "the number of transitions between pixels"; this means that if $d = 1$ and $\theta = 0$, and the algorithm is at pixel $i = I_g(1,1)$, then $j = I_g(1,2)$, the immediate neighbor at right. Suppose that $I_g(1,1) = 255$ and $I_g(1,2) = 128$, then $q(255,128)$ is incremented by 1.

Several descriptors can be extracted from the GLCM addressing contrast (i.e., the amount of local variations) and orderliness (i.e., the regularity of pixel values within an image). We consider two descriptors derived from the GLCM: the angular second moment (ASM), which describes textural homogeneity/uniformity and the entropy, which is proportional to the heterogeneity/randomness:

$$\mathrm{ASM} = \sum_{i,j} (q_{i,j})^2 \tag{3}$$

$$\mathrm{Entropy} = -\sum_{i,j} q_{i,j}(\log q_{i,j}) \tag{4}$$

Using textural descriptors, parametrized by the nearest neighbor pixels, and an isotropic GLCM,[27] we calculated coalescence indicators for STEM of PMO films. The films dominated by spherical pores (58% porous) presented higher textural heterogeneity and lower uniformity (low ASM, high entropy) than those in which the pores coalesced (73% porous). The increase in ASM and decrease in randomness from the "Ordered" to the "Coalesced" sample indicate the disappearance of pore walls as the system goes through the order–disorder transition. The "learning" associated to our approach relies on maximizing class (ordered/disordered) distance using regression over the textural image descriptors.

Through textural analysis, we are able to identify different pore structures that are difficult for humans to detect and confirm the order–disorder transition. Notice that our method focuses on the properties of the micrograph (image), and not necessarily on the roughness/texture of the material itself (specimen). For the latter approach, we refer the reader to Knezevic et al.,[28] who discuss the relationship between texture and changes in crystal lattice orientation, as well as the concept of texture intrinsically connected to the properties of the material.

HIGH THROUGHPUT MICROSCOPY

This section overviews common issues in multiple domains, including data representation techniques to alleviate computation and improve querying over overwhelming data size/rates (e.g., 300 TB/day for light sources).

Data Representation

Image analyses often require maintaining several data copies and at different resolutions, e.g., much of the processing starts with the identification of the sample bulk parts for later inspection of fine structures. It is common to have multiscale representations repeatedly computed and stored in the file system, lacking proper connection to the original data and/or connection to metadata associated to the experiment.

Motivated by these demands, we calculate tiled multi-resolution pyramids at four different scales and store them in HDF5 chunked arrays through BigDataViewer.[29] This plugin enables the user to inspect the data efficiently, compress files (average data size compression rate of 10× with microCT,[30]) encapsulate terabyte-size image datasets, including metadata, and optimize access to multiple scales of

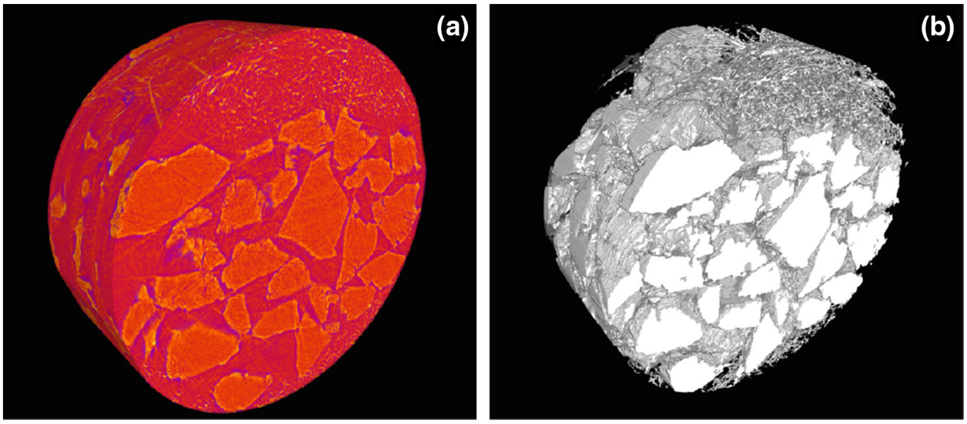


Fig. 5. Segmentation of microCT data, obtained from geological samples, processed using the Parallel Markov Random Field algorithm: (a) rendering original microCT stack and (b) structures after image segmentation.

the data. We build upon Pietzsch's work[31] to explore multi-resolution pyramids and provide a viable way to store and process large image stacks, particularly those illustrated in Figs. 1, 2, and 5.

Searchable Images

Processing semi-structured data created independently into an integrated collection with relevant metadata relies on domain experts driving the annotation process. Compliance with data management and curation policies happens mostly at the level of home institutions or are tailored to a particular experiment. Human interaction has been the main analytical engine for data curation, and this is a challenge to high data throughput instruments and a serious impediment to exascale applications. Research needed includes content-based image retrieval (CBIR) systems to allow fast image queries through image descriptors, potentially leveraging recorded manual interaction and metadata.

In order to enable CBIRs, we have designed several descriptors to detect image singularities from contours. One example is our Multiscale Corner Detection (MCD)[32] method, which applies Ricker wavelet decomposition of the angulation signal to identify saliency points on a shape contour. Our approach assumes that only peaks persisting throughout scales correspond to significant points. MCD detects changes in non-stationary angulation signals, and can be efficiently extended to multidimensional approaches when approximating this wavelet by a difference of Gaussians. Our algorithms explore different scales through correlations, retaining only relevant points of the decomposed angulation signal, and supporting both image compression and signature assemblage. MCD-oriented features have been tested for the purpose of image-based queries, so that visual and semantic content can be properly mined through CBIR systems. Recent work using convolutional neural networks as part of CBIR has proven useful for image retrieval across science domains.[33] To illustrate how these techniques can bring the "faces of scientific data" to reality, we itemize a few mining scenarios with the use-cases from "Across Length Scales" section: (1) describe fiber cross-sections with MCD for later recovery of types of fibers, e.g., unidirectional, which are shapes expected from fibers perpendicular and 45° to the laser beam; (2) retrieve a set of NP images that presents similar NP distribution, according to a certain feature subset; and (3) search thin films that follow ordered pore architecture, given textural descriptors.

PATTERN RECOGNITION

During image partitioning, different phases of a material are identified, which may require billions of voxels to be inspected and grouped together according to criteria varying from local intensity similarity to global spatial organizations. We have developed algorithms that are capable of working on heterogeneous materials, with multiphase/multiregion structures. Investigated methods in data partitioning are PDE-based methods,[34] and graph-based algorithms.[35–37] These algorithms determine underlying structures and provide more efficient, inherently sparse, scale-appropriate data representations of regions of interest (ROI).

This requirement has led us to develop a new scalable image segmentation technique: a Parallel Markov Random Field[38] (PMRF) method for efficient partitioning of large graphs. The Linear and Parallel (LAP) algorithm[37] tailors the MRF model decomposition, and drastically reduces the computational complexity by applying the optimization separately for each subgraph, and hence becomes exponential in the size of the largest graph clique instead. Fig. 5 demonstrates the accuracy of the method, allied to improved processing speed of 26X when compared to its original non-parallel version.[37]

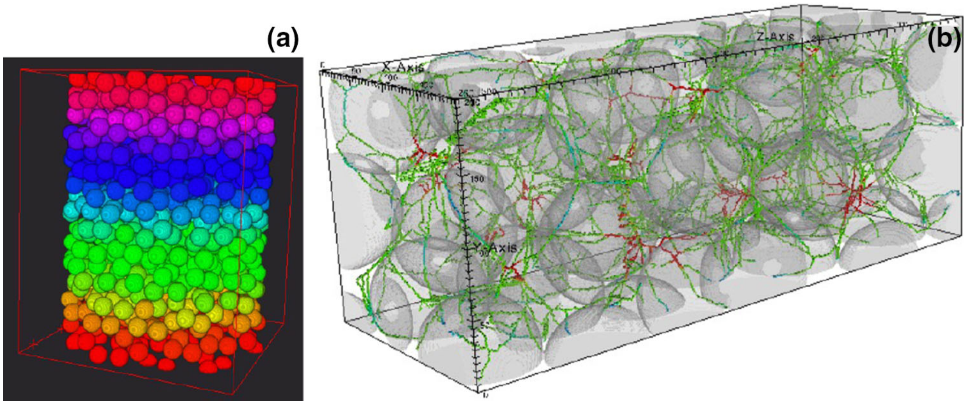


Fig. 6. Simulation of porous material using molecular dynamics code to mimic monodispersive composites and calculation of pore channels: (a) bead-bed color-coded by region; (b) calculated pores and channels from composite.

The transformation of relevant descriptors into decision depends upon sample modeling, domain expertise and ML algorithms; these are instrumental in finding motifs for scientific image analysis. The benefits of exploiting ML are twofold: ML is essential to image partitioning, plus it supports mining repositories of scientific data, as discussed in "Searchable Images" section.

In addition to partitioning/classifying voxels, a higher level data representation, such as a collection of measurements, must be used to understand scientific images, and provide indexing mechanisms for CBIR systems. Compact feature vectors are the result of transforming ROIs into signatures that depict experiments with terabytes of data. Soon, our ML algorithms, allied to our description sets, will allow predictive evaluations through similarity comparison between different experimental samples in terms of much smaller descriptors. A subset of these descriptors includes our work on saliency points from boundary,[32] orientation from texture,[39] and connected networks from topological descriptors.[40] For this reason, we have worked on developing essential elements in data representation,[33] multimodal registration methods[41] and structural classification algorithms for image searching (Fig. 6).

EVOLVING ARCHITECTURES

Building practical computational tools to meet the data explosion has many challenges. One response to the increasing data acquisition rates is to co-locate some computational infrastructure close to the experiment, where new analysis algorithms will be deployed to keep pace with growing data rates. Another response is to exploit evolving computer architectures, which demands re-thinking even the most basic image analysis methods. Appropriate software design will require systematic examination of algorithms in conjunction to the ecosystem where the experiments will be analyzed and stored.

Our F3D plugin (Fig. 2) is an example of how to tackle big data processing by exploring new hardware and specific software designing. Besides using graphics cards technology, F3D provides the capability of applying consecutive accelerated image processing operations to 3D images keeping data in memory from the end-to-end of the complete filtering pipeline, allocating all the necessary memory ahead of the execution, replicating the pipeline across devices and partitioning the data to stream smaller datasets to each device. In doing so, it increases efficiency and decreases the cost of data movement.

Another example is our work on the parallel Markov Random Field (Fig. 5) method[35] that uses a threaded, shared-memory approach, and future versions will explore new distributed-memory schemes such as message passing interface (MPI) and do-it-yourself analysis (DIY).[42] Because of the parallel nature of the PMRF algorithm, chunks of the original large image dataset can be processed separately in different computing cores with minimum communication. In doing so, we expect to take advantage of multi-core/many-core architectures available at scientific computing facilities. We are also investigating other architectures than Von-Neumann, for example, cognitive computing for low-power consumption chips, to deploy ML algorithms such as the convolutional neural networks.

DISCUSSION

Imaging facilities may collect data from materials ranging from bones to geological samples in order to measure only porosity. Also, the science questions from different users about the same sample can widely vary. Nonetheless, these experiments require very similar processing steps before the image content can be used for material quantitation, quality control and decision making.

Algorithms and software to exploit information buried in massive datasets in order to provide knowledge from scientific datasets will have a major impact on experimental science. In order to accomplish this task, we have provided a wide range of algorithms to turn high-resolution multidimensional images into partitioned regions, for the extraction of image descriptors, recovery of microstructures, and classification of samples. These have been wrapped into interactive and customizable scientific frameworks.

This paper introduced the project IDEAL: Images across Domains, Experiments, Algorithms and Learning, and its current software apps to address scientific questions dependent on data collected from imaging instruments across different scales. It also overviewed the analysis algorithms and tools necessary to transform raw data into actionable insights. We discussed image collections from different materials in “Across Length Scales” section, with scales varying from 0.65 to 0.00164 μm, and diverse representation (3D + time, 3D and 2D), which present several commonalities. Examples of these common tasks include: (1) metrics, such as texture analysis, which are shared among imaging modalities; (2) algorithms, such as F3D enhancers and ML methods for data partitioning that can and have been applied across different types of datasets, both for segmentation (within image domain), and sample clustering and classification (among image samples); (3) visualization techniques, such as volume rendering and isocontours, that are performed similarly by different projects; and (4) features and structure detectors that have been used across image domains, for example, blob-detecting descriptors.

Projects dealing with different imaging modalities can share tools, but, when these modalities regard the same sample source, analyses can go even beyond; in these cases, methods to fuse different modalities are essential to discover information that lies on the confluence of heterogeneous imaging sources. Data representation, described in “Data Representation” section, also plays a major role in registration of multimodal data, since saliency points can work as fiducial marks to align images properly and speed up image retrieval (“Searchable Images” section). These algorithmic developments, in concert with evolving architectures (“Evolving Architectures” section), have allowed for scalability of key algorithms to perform pattern recognition.

CONCLUSION

Strategies have been discussed separately, but they are profoundly inter-related and inter-dependent, and this synergy is essential to advance multivariate pattern analysis. We discussed promising descriptors/signatures (e.g., MCD) and ML algorithms (e.g., PMRF), and how they can be used in image-centric problems.

Synergistic aspects of the combination of the projects discussed in previous sections have motivated us to deploy MCD-oriented features to enable image-based querying for CBIR systems. Future developments include exploration of neuromorphic algorithms and efficient schemes to retrieve distributed data. In addition, we will extend our algorithms for image fusion from medical data to material science, for example, to register microCT to SEM.

As a complementary work to the image analysis and pattern recognition, upcoming tools will track recurring computation modes or motifs by considering file-size typical aggregates of scientific datasets, common communication patterns necessary for the analysis, and evaluation of storage demands. These communication patterns will soon prescribe optimized pathways for image analysis at scale.

These algorithms are part of our IDEAL apps that support and accelerate research that requires analyzing information hidden in digital images. Examples of science domains impacted include:

- Crack detection and microdamage evaluation of materials under deformation; new designed composites to be used in construction of jet engines;
- Neuromorphic computing and convolution neural networks applied to problems in which material properties are not well specified and/or computing efficiency is key to embed processing in instruments;
- Quantification of porous material to detect relevant paths and clogging, as part of geological processes involved in carbon sequestration and oil recovery;
- Analysis of geological samples before, during and after fracking in order to quantify environmental impact;
- Development of hierarchically porous materials with prescribed architectures for advanced energy devices;
- Molecule and cell counting, including detection of cell nano-structures with unknown functionality that may play a major role in mechanical regulation and communication intra- and intercell, with application to artificial photosynthesis and the search for biofuels.

ACKNOWLEDGEMENTS

This work was supported by the Director, Office of Science, Advanced Scientific Computing Research and Basic Energy Sciences, of the US Department of Energy. Both the Early Career Research project and the Center for Applied Mathematics for Energy Related Applications (CAMERA) are under Contract No. DE-AC02- 05CH11231. This research used resources of the National Energy Research Scientific Computing Center, which is supported by the Office of Science of the US Department of Energy under Contract No. DE-AC02- 05CH11231. We would like to thank Przemyslaw Oberbek, Materials Design Division, Faculty of Materials Science and Engineering, Warsaw University of Technology, Warsaw, Poland for preparing the SEM images. Also, the authors thank P. Nico and A. Wills for sharing samples of microCT of geological materials, and STEM of PMO, respectively. Additional thanks to B. Loring for supporting visualization schemes, and M. Alegro for participating on the development of multimodal registration methods.

REFERENCES

1. D. Martin, C. Fowlkes, D. Tal, and J. Malik, in *Proceedings of IEEE International Conference on Computer Vision* (2001), p. 416.
2. B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik, in *Proceedings of IEEE ECCV– European Conference on Computer Vision* (2014) p. 297.
3. K. Singha, S. Maity, M. Singha, and S. Pal, *Front. Sci.* 11 (2012).
4. D. Ushizima, A. Bianchi, C. deBianchi, and W. Bethel, in *ImageJ User and Developer Conference* (2012).
5. L. Martin, A. Tuysuzoglu, W.C. Karl, and P. Ishwar, *IEEE Trans. Image Process.* 24(11), 4069 (2015).
6. E.J. Tuegel, A.R. Ingraffea, T.G. Eason, and S. Spottswood, *Int. J. Aerosp. Eng.* 1, 5 (2011).
7. D. Ushizima, T. Perciano, H. Krishnan, B. Loring, H. Bale, D. Parkinson, and J. Sethian, in *Proceedings of IEEE International Conference on Big Data* (2014).
8. A. Krizhevsky, I. Sutskever, and G.E. Hinton, *Adv. Neural Inf. Process. Syst.* 2, 1106 (2012).
9. O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A.C. Berg, and L. Fei-Fei, Int. *J. Comput. Vis.* 115(3), 211 (2015).
10. M. Jordan, *IEEE Spectr.* 1109, (2014). http://spectrum.ieee.org/robotics/artificial-intelligence/machinelearning-maestro-michael-jordan-on-the-delusions-of-big-data-and-other-huge-engineering-efforts.
11. Argonne National Laboratory, TomoPy. https://tomopy.readthedocs.io/en/latest. Accessed 7 July 2016.
12. T. Bicer, D. Gürsoy, R. Kettimuthu, F. De Carlo, and I. Foster, *J. Synchrotron Radiat.* 23, 4 (2016).
13. Argonne National Laboratory, Advanced photon source: an office of science national user facility. https://www1.aps.anl.gov/Science/Scientific-Software. Accessed 7 July 2016.
14. B.N. Cox, H.A. Bale, M. Blacklock, M.N.T. Fast, V. Rajan, R. Rinaldi, R.O. Ritchie, M. Rossol, J. Shaw, Q.D. Yang, F. Zok, and D.B. Marshall, *Annu. Rev. Mater. Res.* 44, 479 (2014).
15. General Electric, Ceramic matrix composites improve engine efficiency. http://www.geglobalresearch.com/innovation/ 2016. Accessed 7 July 2016.
16. H.A. Bale, A. Haboub, A.A. Macdowell, J.R. Nasiatka, D.Y. Parkinson, B.N. Cox, and D.B. Marshall, *Nat. Mater.* 12, 40 (2012).
17. R. Nock and F. Nielsen, *IEEE Trans. Pattern Anal. Mach. Intell.* 26, 1452 (2004).
18. Fiji, Imagej. http://pacific.mpi-cbg.de/wiki/index.php/Fiji. Accessed 7 July 2016.
19. R.C. Gonzalez and R.E. Woods, *Digital Image Processing*, 3rd edn. (Prentice-Hall, 2006), p. 861.
20. R.O. Duda, P. Hart, and D.G. Stork, *Pattern Classification* (Wiley, New York, 2001).
21. A. Gajewicz, N. Schaeublin, B. Rasulev, E. Maurer, S. Hussain, T. Puzyn, and J. Leszczynski, *Nanotoxicology* 9, 313 (2014).
22. A. Mikolajczyk, A. Gajewicz, B. Rasulev, N. Schaeublin, E. Maurer-Gardner, S. Hussain, J. Leszczynski, and T. Puzyn, *Chem. Mater.* 27, 2400 (2015).
23. E. Burello and A.P. Worth, *Wiley Interdiscip. Rev. Nanomed. Nanobiotechnol.* 3(3), 298 (2011).
24. K. Odziomek, D. Ushizima, M. Haranczyk, and T. Puzyn, in *Proceedings of American Chemical Society* (2014).
25. A.W. Wills, D.J. Michalak, P. Ercius, E.R. Rosenberg, T. Perciano, D. Ushizima, R. Runser, and B.A. Helms, *Adv. Funct. Mater.*, 1 (2015).
26. D.M. Ushizima-Sabino, L. da Fontoura Costa, E.G., Rizzatti, and M.A. Zago, *Real Time Imaging* 10(4), 205 (2004).
27. Y. Zhong, and A.K. Jain, *Pattern Recogn.* 33, 671 (2000).
28. M. Knezevic, A. Levinson, R. Harris, R.K. Mishra, R.D. Doherty, and S.R. Kalidindi, *Acta Mater.* 58(19), 6230 (2010).
29. T. Pietzsch, S. Saalfeld, S. Preibisch, and P. Tomancak, *Nat. Methods* 12, 6 (2015).
30. D. Ushizima, T. Perciano, and D. Parkinson, in *Proceedings of IEEE International Conference on Big Data* (2014).
31. T. Pietzsch, in *BigDataViewer.*http://fiji.sc/BigDataViewer. Accessed 7 July 2016.
32. I. Paula Jr., F. Medeiros, F. Bezerra, and D. Ushizima, *J. Math. Imaging Vis.* 45, 251 (2013).
33. F. Araujo, R. Silva, and D.M. Ushizima, in *Proceedings of PyData San Francisco* (2016).
34. J. Sethian, *Level Set Methods and Fast Marching Methods: Evolving Interfaces in Computational Geometry, Fluid Mechanics, Computer Vision, and Materials Science* (Cambridge University Press, Cambridge, 1999).
35. T. Perciano, D. Ushizima, E.W. Bethel, Y.D. Mizhahi, and J.A. Sethian, in *Proc.eedings of IEEE ICIP Conference* (2016).
36. D. Ushizima, A. Bianchi, and C. Carneiro, in *Proceedings of IEEE ISBI Symposium* (2014).
37. Y.D. Mizrahi, M. Denil, and N. de Freitas, *Proc. ICML* 32, 1 (2014).
38. S.Z. Li, *Markov Random Field Modeling in Image Analysis* (Springer, London, 2009).
39. G.C. Leite, D.M. Ushizima, F.N.S. Medeiros, and G.G. de Lima, *Sensors* 10(6), 5994 (2010).
40. D.M. Ushizima, D. Morozov, G.H. Weber, A.G. Bianchi, J.A. Sethian, and E.W. Bethel, *IEEE. Trans, Vis. Comput. Gr.* 18(12), 2041 (2012).
41. M. Alegro, E. Amaro-Jr, B. Loring, H. Heinsen, E. Alho, L. Zollei, D. Ushizima, and L.T. Grinberg, in *Proceedings of IEEE CVPR Conference* (2016).
42. T. Peterka, R. Ross, A. Gyulassy, V. Pascucci, W. Kendall, H.W. Shen, T.Y. Lee, and A. Chaudhuri, in *Proceedings of IEEE LDAV Symposium* (2011) p. 105.