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, DM
Bale, HA
Bethel, EW
et al.

Publication Date
2016-11-01

DOI
10.1007/s11837-016-2098-4

Peer reviewed
SAIDE: Scaling Analytics for Image-based Data from Experiments *

Daniela Ushizima1,6, Hrishikes Bale2, Wes Bethel1, Peter Ercius3, Brett Helms2, Harinarayan Krishnam1, Lea Grinberg3, Alastair Macdowell4, Katarzyna Odziomek5, Dula Parkinson4, Talita Perciano1, Robert Ritchie2,6, Chao Yang1, James Sethian1,6

1Computational Research Division, Lawrence Berkeley National Laboratory
2Materials Science Division, Lawrence Berkeley National Laboratory
3Memory and Aging Center, University of California San Francisco
4Advanced Light Source Division, Lawrence Berkeley National Laboratory
5Chemistry Department, University of Gdansk, Poland
6University of California Berkeley

ABSTRACT
Research across science domains is increasingly reliant on image-based data from experiments. The challenge is to analyze the data torrent generated by advanced instruments in a timely manner and provide insights such as measurements for decision-making. Software tools are in high demand for scientists to uncover relevant, but hidden, information in digital images, such as those coming from new materials. A group of computational and material scientists have embraced the multi-disciplinary work of designing software applications, coordinating research efforts connecting (1) emerging algorithms for dealing with complex and large datasets; (2) data analysis methods with basis on pattern recognition and machine learning; and (3) advances in evolving computer architectures. These new trends will accelerate the analyses of image-based recordings, scaling scientific procedures by reducing time between experiments, increasing efficiency, and opening more opportunities for more users of the imaging facilities. This paper aims to provide an overview of our algorithms and deployed software tools, showing results across image scales and how each tool component plays a role in improving image understanding within DOE national laboratories.

1. INTRODUCTION
Emerging technologies in CMOS and CCD image sensors have made digital images ubiquitous. Collecting experimental outcomes by keeping image records of relevant evidences has become a common practice, but automated algorithms for information extraction of such datasets have not evolved at the same pace. The increasing rate at which images can be produced, combined with the broad variety of sensors, poses algorithmic challenges when applied to large volumes of data, particularly when it contains heterogeneous structures in multiple scales.

To characterize microstructures, a material scientist analyzes samples at different angles, granularities, and at varying mechanical conditions. With the ability of acquiring micrographs at high spatial and temporal resolution, 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. But in so doing, the precision and subtlety – the information obtained by expensive and high resolution instruments – vanish.

These challenges are compounded by the lack of a cohesive strategy that fuses data understanding, algorithms, computing hardware advances, and high bandwidth data transfer. Detectors will create much more data, and rapidly. Network advances will allow us to transfer data to far away computing facilities. Advances in computer architectures will provide orders of magnitude processing speedup. To take advantage of these scientific progresses, we have constructed coherent, scalable and novel computing approaches that take five key components into consideration: detectors, images, algorithms, data representation and computing architectures.

We can 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 manual-labor demanding, one-off workflows often include thresholding and hand-tuned filters applied to individual images, followed by measurements.

A second possible path is to adopt more sophisticated, yet generic and fully automated pipelines that find regions of interest by calculating subspace partitions [1, 2] from large image sets. Also fast to deploy, but involving less manual interaction, these pipelines can present the following issues: (a) work only on simulated data [3, 4] or (b) the segmentation inaccuracies undermine measurements at subsequent processing tasks, such as structure assessment [5].
Finally, we have pursued a third strategy, combining relevant tasks from the previous categories to account for the high data throughput regime (over terabytes per experiment) while imposing controlled amounts of human labor. By tracking manual interaction while working with one-off prototypes and generic tools, we record user interaction and mine collected data from user to generate models than can both provide simulated data [6] and information to create motifs for scientific image analysis.

The key impacts of this work are:
- Extract information from noisy data, including modeling of artifacts inherent to the detector;
- Develop geometrical descriptions of materials, in particular create analyzable 3D models from 2D scans that allows data compression, while still retaining essential content;
- Initialize numerical models and establish check points for numerical simulations to verify if models match experimental data;

These algorithms are part of tools to support and accelerate DOE research that requires analyzing information hidden in digital images. Examples of science domains impacted:
- 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 descriptors are not well specified and/or computing efficient is key to embed processing to instruments or close to data collection;
- 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 fracking in order to quantify environmental impact;
- Development of new architectures using periodic mesoporous organisilica for designing the next generation, more efficient computer chips;
- 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 inter-cell, with application to artificial photosynthesis and the search for biofuels.

This paper describes our main algorithms and summarizes important discoveries in material science, involving ceramic composites, polymeric films and nanoparticles.

### Table 1: Material microscopy images across scales: science problems.

<table>
<thead>
<tr>
<th>Materials</th>
<th>Resolution (µm)</th>
<th>Image modality</th>
<th>Physics at pixel/voxel</th>
<th>Data analysis for specimen quantification</th>
</tr>
</thead>
<tbody>
<tr>
<td>fibrous composites</td>
<td>0.65 to 1.3</td>
<td>microCT</td>
<td>X-ray attenuation contrast</td>
<td>detection of fibers, fiber breaks and cracks; structural dynamics translated into similarity index variations.</td>
</tr>
<tr>
<td>geological samples</td>
<td>0.65 to 2.5</td>
<td>microCT</td>
<td>X-ray attenuation contrast</td>
<td>segregation of components from multiphase specimens; analysis of homogeneity for parts; pockets, pore network and global orientation of channels.</td>
</tr>
<tr>
<td>nanoparticles</td>
<td>0.3267</td>
<td>SEM</td>
<td>electron absorption</td>
<td>counting, topographical characterization, morphology and particle distribution; structural dynamics translated into similarity index variations.</td>
</tr>
<tr>
<td>thin films</td>
<td>0.00164</td>
<td>STEM tomography</td>
<td>electron transmission</td>
<td>pore organization across film, associated to level of pore coalescence; surface density analysis correlated to dielectric constant measurements.</td>
</tr>
</tbody>
</table>

### 2. BACKGROUND

Although there is an abundance of stories about how computers tackle vision problems [7, 8], applying machine vision to the understanding of scientific image analysis remains mostly unsolved.

ImageNet [7] offers tens of millions of cleanly sorted images to train classifiers, but experimental data presents surprisingly variable images, even when considering the same instrument. For example, the LBNL Advanced Light Source beamline 8.3.2 receives samples ranging from bones to geological samples in order to measure osteopenia and calcium precipitation, respectively. Also, the science questions from different users about the same sample can widely vary.

What to do when it is impossible to collect millions of images? According to Krizhevsky et al [8], datasets of labeled images on the order of tens of thousands of images are relatively small and allow for simple recognition tasks.

Although imaging facilities can generate petabytes of data, image collections are far from being cleanly sorted. Nevertheless, we are confident that variability can be constrained in many cases, particularly when modeled and cleanly sorted and labelled through image simulators. In these cases, synthetic samples carry fundamental characteristics that well-represent the problem domain (augmented with label-preserving transformations).

### 3. IMAGE ACROSS SCALES

Definition, processing, compression, and identification of material properties using computer vision algorithms have enabled quantitative data analyses, previously requiring laborious manual interaction within experimental facilities. By selecting specific image problems to balance depth and breadth in engineering science domain applications, we have deployed key methods across scales to enable: (i) Analysis of ceramic composites before constructing jet engine for the assessment of material fracture; (ii) Quality control of polymeric films with application to microelectronics to minimize dielectric constant for mesoporous organonics composites, and (iii) Standardization of experimental records using nanoparticle microscopy as input to morphometrical algorithms, which correlate toxicity to nanoparticle morphology. These different materials, and respective analysis algorithms, are the topics of the next sections.

#### 3.1 Ceramic matrix composites
Ceramic matrix composites or CMCs provide exceptional strength-to-weight ratio capabilities, appropriate to the construction of jet engines. For such a purpose, exposure to high temperature and microstructural mechanisms may require metal replacement by reinforced CMCs [7, 9]. SiC-based composites are a type of CMC, and the samples described in this paper were fabricated at Hypertherm (Huntington Beach) by winding SiC fibers (Figure 1) on a carbon frame [10]. The goal is to monitor material resistance for high performance design, with a broad array of applications at LBL, UCB, Air Force, General Electric and NASA.

Tenacity tests often involve subjecting samples to varying forces and different temperatures during 3D data acquisition using hard X-ray. The resulting micro-tomography images (micro-CT) represent a 3D map of the X-ray attenuation coefficient of the materials within the sample. Inspection involves three key factors: (a) identification of the different components of the sample; (b) detection of microcracks associated to specific conditions of strain and temperature; (c) location of fibers and pull-outs within the material [10].

The fibers present a typical 2D-profile, which we use as input to a template matching algorithm. Initially, the user quickly selects a few examples or prototypes (square patches) that represent the expected fiber profiles, next our algorithm runs two main calculations several times: the mean square error (MSE) and the normalized cross-correlation coefficient (NCC). The MSE determine the most suitable prototype $p$ and the NCC computes the similarity between each pixel of the image and $p$.

The MSE and CC are defined as:

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

and

\[ CC(x, y) = \frac{\sum_{i,j} (p_{ij} - \bar{p}) \sum_{i,j} (f_{ij} - \bar{f})}{\sqrt{\sum_{i,j} (p_{ij} - \bar{p})^2 \sum_{i,j} (f_{ij} - \bar{f})^2}} \]

where $i, j$ are image indexes, $f_{ij} = f(x + i, y + j)$ for simplicity, $\bar{\cdot}$ 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 $[-1, 1]$.

We have constructed image processing tools for users of CMC microCT data in order to provide (near) real-time feedback of post processed or in-situ processed data. As an example, we have designed F3D [11], a platform portable library that ensures algorithms are usable to a wide range of users. We deliver image processing algorithms necessary to accelerate analysis of GB/TB scale data with our GPU-aware 3D fast filtering package (F3D). Fig.2 shows the interface of our tool, which contains key algorithms to enhance structures of interest for detection of fiber breaks and ceramic matrix cracks, and are crucial to microstructure characterization.

### 3.2 Films for microelectronics

Controlling a material porosity is 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 enables assembly of well-controlled, ordered architectures. This section describes some of the factors governing porogen packing and shape persistence during mesoscale assembly.
In [12], we describe image attributes 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 domains in 3D space. Our contribution is to define a set of statistical descriptors for STEM images that indicates pore packing relationships and pore organization across 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 [13]. 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 \( q \) of an image \( I_q \) at every two pixels, \( i = I_q(x, y) \) and \( j = I_q(\hat{x}, \hat{y}) \), given a direction \( \theta \) and distance \( d \). In other words, we calculate the Gray-Level Co-occurrence Matrix (GLCM) to measure the spatial organization of the pixel intensities, using the following equation:

\[
\begin{align*}
q(i, j|d, \theta) &= \#(i, j) \in I_q|j = \rho(i|d, \theta), \\
\rho(i|d, \theta) &\rightarrow \hat{x} = x + d_1, \hat{y} = y + d_2,
\end{align*}
\]

where \( \theta \in 0, 45, 90, 135 \) and \( d_2 \in [-1, 0, 1] \). 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: angular second moment (ASM), which describes textural homogeneity/uniformity and entropy, which is proportional to the heterogeneity/randomness

\[
ASM = \sum_{i,j} (q_{i,j})^2, \\
Entropy = -\sum_{i,j} q_{i,j} \log q_{i,j}
\]

Using textural descriptors, parametrized by the nearest neighbor pixels, and isotropic direction, we calculate coalescence indicators for STEM of PMO films. The films dominated by spherical pores (58% porous) presented higher textural heterogeneity and lower textural uniformity (high low ASM) than those in which the pores coalesced (73% porous). In a porous film, the GLCM descriptors are sensitive to void-wall interfaces. High ASM and low entropy descriptors indicate more interfaces. 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. Textural analysis therefore allows us to separate different pore structures in images with complex structures that are difficult for the human eye to assign and provides further confirmation of an order-disorder transition.

### 3.3 Toxicity of nanoparticles

When people look at scanning electron (SEM) images, they usually focus at the nanoparticle size distribution. Nevertheless, a lot more information can be extracted, e.g., particle shape and roughness [?]. These features are essential because the way some nanoparticles interact is heavily reliant on their morphology or surface structure. The features that we extracted from the SEM images are incorporated into models that guide compression of an image ensemble to a minimum representative set. This processing allows laboratories to retain the nanoparticle population statistics while selecting only the image files that represent the ensemble distribution to be kept on record.

Fig. 4 shows tricalcium phosphate \( \text{Ca}_3(\text{PO}_4)_2 \), a naturally occurring mineral with a wide range of biomedical applications, containing a distribution of nanoparticles of various shapes and sizes. One of the challenges of working with nanoparticles is the fact that they do not act like bulk molecules and they do not act like single molecules they are something in between, so conventional modeling tools have limited usage. We have designed new tools to...

### 4. HIGH THROUGHPUT MICROSCOPY

Digital images, as part of experimental records, are acquired in research laboratories in all science domains. Overwhelming data size/rates (e.g. 300TB/day for light sources) make storage an issue. Invaluable information, encoded in high-resolution images and obtained at considerable cost, was at risk. Several users depend upon tools that require drastic downsampling in order to analyze overwhelming data size/rate. DOE identified and reported several of these bottlenecks [14, 15], emphasizing that analysis of experimental data coming from DOE high-throughput sensors is fundamental to applied science. We have addressed several of these challenges by deploying tools that take advantage of...

---

**Figure 3:** Peter: please description.

**Figure 4:** Nanoparticle detection from SEM, quantification and characterization of ensembles.
Figure 5: Segmentation of microCT image of geological sample using the Parallel Markov Random Field approach: (a) Volume rendering of the original microCT stack (b) Volume rendering of the structures after segmentation.

Figure 6: Simulation of porous material using molecular dynamics code to mimic monodispersive composites and calculation of pore channels.

data presentation methods and pattern recognition algorithms.

4.1 Data representation

5. PATTERN RECOGNITION

Without doubt, generic and public tools for image analysis are available, but seldom built so that they can scale to large data sets and explore heterogeneous architectures. The usefulness of current methods is often restricted to one of the stages of processing, are too generic to exploit the particulars of the acquisition systems, and/or are inaccessible to all but the most specialized image analysts.

Understanding the changes requires not just visualizing the data, but identifying and quantifying features of interest and tracking them in time. Our task is to automate this analysis for the huge amounts of data. Approaches include separating the phases that compose the materials using a parallel Markov Random Field method (PMRF) [16] and pattern matching to detect and measure specific features in the samples. Original results shown in Fig. 5 demonstrate the accuracy of the method, allied to improved processing speed of 26x when compared to its original non-parallel version of the method [17].

5.1 Data curation

digital twin, simulation, andreas’ tools

5.2 Data analysis and understanding

In order to go from experimental images to volumes of interest, we have deployed image enhancement and partitioning methods, with filtering and segmentation schemes to detect micro-structures from large data sizes in a reasonable amount of time. Our algorithms are capable of working on heterogeneous materials, with multiphase/multi-region structures. We have explored a host of emerging technologies in data partitioning, such as PDE-based methods [18], and graph-based algorithms [7, 19, 17]. These algorithms determine underlying structures and provide more efficient, inherently sparse, scale-appropriate data representations of regions of interest (ROI).

For example, we have just deployed a new scalable image segmentation using Markov Random Fields (MRF) [20] methods and efficient graph partitioning. This approach allows parallel MRF optimization and parameter estimation for very large graphs. The Linear and Parallel (LAP) algorithm [17] tailors the MRF model decomposition, such as it reduces the computational complexity drastically by applying the optimization separately for each subgraph, being exponential in the size of the largest graph clique instead.

Different science domains bring specific descriptors, which we have translated to compact feature vectors by transforming ROIs into signatures that depict experiments with terabytes of data. We have worked on representations to handle large 3D still images and sequences, and soon 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 [21], orientation from texture [22], and connected networks from topological descriptors [23]. These are essential elements in both multimodal registration methods and structural classification algorithms.

The transformation of relevant descriptors into decision relies on sample modeling, domain expertise and machine learning - these are instrumental in finding motifs for scientific image analysis.

The benefits of exploiting ML are two-fold: ML is essential to image partitioning, plus it can also be applied to understanding other data such as text-centric datasets, e.g., in text-mining repositories of scientific documents [24].

6. EVOLVING ARCHITECTURES

Building practical computational and software tools to meet the data explosion has many challenges. One response to 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 will require re-thinking even the most basic image analysis algorithms. Appropriate software design will require investigation of algorithms in conjunction to the ecosystem where the experiments will be analyzed and stored.

Shared-mem and distributed-mem to cite Perciano Colleen:2016 - to be submitted

Chao: a sentence about neuromorphic computing

Lea: a sentence about multimodal registration

7. CONCLUSION

8. DISCUSSION
What is the impact? Algorithms and software to exploit information buried in massive datasets in order to transform raw experimental data and simulations into knowledge and actionable insights, which will have a major impact on experimental science. In order to accomplish this task, we will provide a wide range of algorithms to prepare high resolution multidimensional images to be partitioned into regions of interest, extract data descriptors, recover microstructures, and classify samples. These will be wrapped into interactive and customizable scientific frameworks that benefit from a high performance infrastructure. These tasks will be monitored on the HPC facility side to provide more efficient use of DOE resources.

Why image analysis?

Why machine learning?

What is critical? Partitioning methods that can mathematically be extended to any dimension and that can change topology naturally are essential in multivariate pattern analysis. Also, the best data decomposition is the one that leads to the sparsest representation. We will work on methods that explore sparse representations as wavelets, curves and compressive sensing for data representation, feature extraction and spatial data compression.

How does other areas benefit? As a complementary work to the image analysis and pattern recognition, we have tracked recurring computation modes or motifs [25] by looking into 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.

9. ADDITIONAL AUTHORS

10. REFERENCES


