Suspension Crystallization Monitoring using In Situ Video Microscopy, Model-based Object Recognition, and Maximum Likelihood Estimation

by

Paul A. Larsen

A PRELIMINARY REPORT

SUBMITTED IN PARTIAL FULFILLMENT OF THE

REQUIREMENTS FOR ADMISSION TO CANDIDACY

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

(Chemical and Biological Engineering)

at the

UNIVERSITY OF WISCONSIN–MADISON

2007

Suspension Crystallization Monitoring using In Situ Video Microscopy, Model-based Object Recognition, and Maximum Likelihood Estimation

Paul A. Larsen
Under the supervision of Professor James B. Rawlings
At the University of Wisconsin–Madison

Video imaging can provide a wealth of information about chemical processes. Extracting that information in a way that enables better process understanding and control typically requires methods tailored to the particular application. For applications involving particulate processes, usually the goal is to measure the size and shape distributions of the particles. In many cases, the particle shape is a marker for the particle's polymorphic form, or internal structure. Measuring the sizes and shapes of the particles requires image segmentation, or separating the objects of interest (i.e. the particles) from the image background. The information obtained by successful segmentation can be useful but can also be misleading unless the imaging measurement is fully understood. Any imaging-based measurement is subject to sampling bias due to the finite size of the imaging frame. Sampling bias may also result due to the nature of the segmentation method employed. The degree to which these biases affect the obtained information depends on both the imaging system and the properties of the particle population.

To address the challenges associated with effective use of video imaging for particulate processes, this thesis focuses on the following areas:

1. Developing image analysis algorithms that enable segmentation of noisy, in situ video images of crystallization processes.

2. Developing statistical estimators to overcome the sampling biases inherent in imaging-based measurement.

3. Characterizing the reliability and feasibility of imaging-based particle size distribution measurement given imperfect image analysis and time constraints.

We have developed two image analysis algorithms. The first algorithm is designed to extract particle size and shape information from in situ images of suspended, high-aspect-ratio crystals. This particular shape class arises frequently in pharmaceutical and specialty chemical

applications and is problematic for conventional monitoring technologies that are based on the assumption that the particles are spherical. The second algorithm is designed to identify crystals having more complicated shapes. The effectiveness of both algorithms is demonstrated using in situ images of crystallization processes and by comparing the algorithm results with results obtained by human operators. The algorithms are sufficiently fast to enable real-time monitoring for typical cooling crystallization processes.

Using a statistical model of the particle imaging process, we have generated artificial images of particle populations under a wide variety of imaging and process conditions, enabling us to explore the biases inherent in the imaging process and develop methods to compensate for such biases. Specifically, we have developed a maximum likelihood estimator to estimate the particle size distribution of needle-like particles. We benchmark the estimator against the conventional Miles-Lantuejoul approach using several case studies, showing that both methods adequately compensate for the finite-frame-size sampling bias. For needle-like particles, our estimator provides better estimates than the Miles-Lantuejoul approach, but the Miles-Lantuejoul approach can be applied to a wider class of shapes. Both methods assume perfect image segmentation, or that every particle appearing in the image is identified correctly.

Given that perfect image segmentation is a reasonable assumption only at low solids concentrations, we have developed a descriptor that correlates with the reliability of the imaging-based measurement (i.e. the quality of the image segmentation) based on the amount of particle overlap. We have demonstrated how both the Miles-Lantuejoul and maximum likelihood approaches discussed above underestimate the number density of particles and have developed a practical approach for estimating the number density of particles for significant particle overlap and imperfect image analysis. The approach is developed for monodisperse particle systems.

Finally, we assess the feasibility of reconstructing a particle size distribution from imaging data given the time constraints imposed by image acquisition limitations and crystallization kinetics.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Crystallization overview

Crystallization plays a critical role in numerous industries for a variety of reasons. In the semiconductor industry, for example, crystallization is used to grow long, cylindrical, single crystals of silicon with a mass of several hundred kilograms. These gigantic crystals, called boules, are sliced into thin wafers upon which integrated circuits are etched. Prior to etching, crystallization is used to grow thin layers of crystalline, semiconductor material onto the silicon wafer using a process called chemical vapor deposition. In the food industry, crystallization is often used to give products the right texture, flavor, and shelf life. Crystallization is used to produce ice cream, frozen dried foods, chewing gum, butter, chocolate, salt, cheese, coffee, and bread [46]. These examples highlight the utility of crystallization in creating solids with desirable and consistent properties.

Crystallization is also widely used to separate and purify chemical species in the commodity, petrochemical, specialty, and fine-chemical industries. In fact, DuPont, one of the world's largest chemical manufacturers, estimated in 1988 [42] that approximately 70% of its products pass through a crystallization or precipitation stage.

Crystallization is a critical process for the pharmaceutical industry. The vast majority of pharmaceuticals are manufactured in solid, generally crystalline, form. Crystallization is used to identify structure for use in drug design, to isolate chemical species from mixtures of reaction products, and to achieve consistent and controlled drug delivery.

Although each of the industries mentioned above is important to the U.S. economy, generating 35% of the United States' gross domestic product in 2003, the semiconductor and pharmaceutical industries are particularly critical to the U.S. economy. Semiconductor Industry Association (SIA) President George Scalise stated in the SIA 2004 Annual Report that "the U.S. economy is hardwired to semiconductor innovation." After suffering severe setbacks in 2001, the semiconductor industry rebounded and is expected to maintain 8 to 10 percent growth in the coming years. According to the SIA, worldwide semiconductor sales in 2004 exceeded USD$213 billion, with the U.S. gaining nearly 47% of the market [98]. A 2004 study conducted by the Milken Institute [73] found that the biopharmaceutical industry directly employed over 400,000 people in 2003 while generating 2.3 million additional jobs, making it responsible for over 2 percent of the total employment in the United States. The study found furthermore that the biopharmaceutical industry

generated USD$115 billion dollars in total workers' earnings in 2003, including USD$29.5 billion direct impact (pharmaceutical company employees), USD$54.3 billion indirect impact (employees of related companies, such as raw material suppliers or vendors), and USD$31.3 billion induced impact (employment as a result of direct and indirect spending).

Given the widespread use of crystallization in these industries, achieving good control over crystallization processes is clearly an economically significant goal. In this article, we explain how crystallization is controlled in industrial processes and what challenges must be overcome to achieve better control.

1.2 Thesis overview

Chapter 2

Literature review

2.1 Crystallization tutorial

Crystallization is the formation of a solid state of matter in which the molecules are arranged in a regular pattern. Crystallization can be carried out by a variety of methods, but the concepts and terminology relevant to most crystallization processes can be understood by examining the method of solution crystallization. In solution crystallization, the physical system consists of one or more solutes dissolved in a solvent. The system can be *undersaturated*, *saturated*, or *supersaturated* with respect to species i, depending on whether the solute concentration c_i is less than, equal to, or greater than the *saturation concentration* c_i^*. Crystallization occurs only if the system is supersaturated. The *supersaturation level* is the amount by which the solute concentration exceeds the saturation concentration, and is commonly expressed as $\sigma = \frac{c_i - c_i^*}{c_i^*}$, $S = \frac{c_i}{c_i^*}$, or $\Delta c = c_i - c_i^*$. The supersaturation level can be increased either by lowering the saturation concentration (for example, by cooling as depicted in Figure S1) or by increasing the solute concentration (by evaporating the solvent, for example).

Crystallization moves a supersaturated solution toward equilibrium by transferring solute molecules from the liquid phase to the solid, crystalline phase. This process is initiated by *nucleation*, which is the birth or initial formation of a crystal. Nucleation occurs, however, only if the necessary activation energy is supplied. A supersaturated solution in which the activation energy is too high for nucleation to occur is called *metastable*. As the supersaturation level increases, the activation energy decreases. Thus spontaneous nucleation, also called *primary nucleation*, occurs only at sufficiently high levels of supersaturation, and the solute concentration at which this nucleation occurs is called the *metastable limit*. Since primary nucleation is difficult to control reliably, primary nucleation is often avoided by injecting crystal *seeds* into the supersaturated solution.

Crystal nuclei and seeds provide a surface for crystal growth to occur. Crystal growth involves solute molecules attaching themselves to the surfaces of the crystal according to the crystalline structure. Crystals suspended in a well-mixed solution can collide with each other or with the crystallizer internals, causing crystal attrition and breakage that results in additional nuclei. Nucleation of this type is called *secondary nucleation*.

The rates at which crystal nucleation and growth occur are functions of the supersaturation level. The goal of crystallizer control is to balance the nucleation and growth rates to achieve

the desired crystal size objective. Often, the size objective is to create large, uniformly sized crystals. Well-controlled crystallization processes operate in the *metastable zone*, between the saturation concentration and the metastable limit, to promote crystal growth while minimizing undesirable nucleation.

2.2 Crystallization process control in industry

The objective of every industrial crystallization process is to create crystals that meet specifications on size, shape, composition, and internal structure. This objective is achieved using a variety of methods and equipment configurations depending on the properties of the chemical system, the end-product specifications, and the production scale. Continuous crystallizers, such as those shown in Figures S2 and S3, are typically used for large-scale production, producing hundreds of tons per day. In the specialty chemical, fine chemical, and pharmaceutical industries, batch crystallizers (see figures S4 and S5) are often used to produce low-volume, high-value-added chemicals.

2.2.1 Process development

The first step in developing a control system for solution crystallization is to determine the saturation concentration and metastable limit of the target species over a range of temperatures, solvent compositions, and pH's. The saturation concentration, also called solubility, represents the minimum solute concentration for which crystal growth can occur. The metastable limit, on the other hand, indicates the concentration above which undesirable spontaneous nucleation occurs (see the "Crystallization tutorial" sidebar). Spontaneous nucleation, which yields smaller, non-uniform crystals, can be avoided by injecting crystal "seeds" into the crystallizer to initialize crystal growth. The saturation concentration and metastable limit provide constraints on the operating conditions of the process and determine the appropriate crystallization method. For example, chemical systems in which the solubility is highly sensitive to temperature are crystallized using cooling, while systems with low solubility temperature dependence employ anti-solvent or evaporation crystallization. Automation tools greatly reduce the amount of time, labor, and material previously required to characterize the solubility and metastable limit, enabling a wide range of conditions to be tested in a parallel fashion [11].

Once a crystallization method and solvents are chosen, kinetic studies are carried out on a larger scale (tens to hundreds of milliliters) to characterize crystal growth and nucleation rates and to develop an operating policy (see Figure S6) that is robust to variations in mixing, seeding, and impurity levels. These studies minimize the difficulty in scaling up the process several orders of magnitude to the pilot scale. The operating policy is usually determined semi-quantitatively, using trial-and-error or statistical-design-of-experiment approaches. Process robustness is achieved by adopting a conservative operating policy at low supersaturation levels that minimize nucleation events and thus achieve larger, more uniform crystals. Operating at low supersaturation

levels, far from the metastable limit, is important because the metastable limit is difficult to characterize and is affected by various process conditions that change upon scaleup, such as the size and type of vessel or impeller.

2.2.2 Controlled, measured, and manipulated variables

The primary concern of most industrial crystallization processes is generating crystals with a particle size distribution (PSD) that enables efficient downstream processing. The controlled variable for most crystallization processes, however, is the supersaturation level, which is only indirectly related to the PSD. The supersaturation level affects the relative rates of nucleation and growth and thus determines the PSD. Because of its dependence on temperature and solution composition, the supersaturation level can be manipulated using various process variables such as the flow rate of the cooling medium to the crystallizer jacket and the flow rate of anti-solvent to the crystallizer.

Process development studies use a wide range of measurement technology. This technology includes, for example, turbidity probes to detect the presence of solid material, laser scattering to characterize particle size distributions, and spectroscopic or absorbance probes to measure solute concentrations. However, large-scale, industrial crystallizers rarely have these advanced measurements available. In fact, controllers for most industrial crystallizers rely primarily on temperature, pressure, and flow rate measurements. For example, Figure S7 illustrates a commonly used cascade control strategy that uses temperature feedback to follow a prespecified temperature trajectory. This control system is clearly not robust to deviations from the temperature trajectory or disturbances affecting the saturation concentration and metastable limit, as depicted in Figure S6(b).

2.2.3 Nonlinearities and constraints

The nonlinearities inherent in crystallization processes complicate controller design. Crystal growth rate is commonly described using the semi-empirical power law

$$G(T) = k_g e^{-E_g/RT}(c_i - c_i^*(T))^g, \tag{2.1}$$

indicating a highly nonlinear temperature dependence. In (2.1), k_g is the growth rate constant, E_g is the activation energy, R is the gas constant, g is the power-law constant, and $c_i - c_i^*(T)$ is the supersaturation level of species i. The variables c_i and c_i^* are, respectively, the solute concentration and saturation concentration of species i. The saturation concentration is usually also a nonlinear function of temperature. The nucleation rate is typically described using a similar, nonlinear expression.

Control systems for industrial crystallizers must also handle several constraints. The rate at which a crystallizer can be heated or cooled is limited by constraints on the temperature and flow rate of the cooling and heating fluids. The use of shape-modifying additives is constrained

by crystal purity requirements. As mentioned previously, crystallizers are constrained to operate in the metastable zone to avoid spontaneous nucleation and achieve uniformly sized crystals.

2.3 Advanced crystallizer control

The above discussion illustrates the limited technology used to control industrial crystallization processes. The obstacles that have hindered the implementation of advanced PSD control in industry, however, are being overcome by recent advances in measurement and computing technology [16],[89].

With improved control technology, additional challenges can be addressed. One challenge is to control shape, which, like PSD, affects the efficiency of downstream processes such as solid-liquid separation, drying, mixing, milling, granulation, and compaction. In some cases, particularly for chemicals having low solubility or low permeability, the crystal size and shape affect product properties such as bioavailability and tablet stability. Chemical purity must also be controlled, especially for food and pharmaceutical products intended for consumption and for semiconductor devices requiring highly consistent properties.

Perhaps the most difficult and important challenge is controlling *polymorphism*, which is the ability of a chemical species to crystallize into different crystal structures. The polymorphic form affects product characteristics, including stability, hygroscopicity, saturation concentration, dissolution rate, and bioavailability. The development of increasingly complex compounds in the pharmaceutical and specialty chemical industries makes polymorphism a commonly observed phenomenon for which control is essential. The recent disaster at Abbott Labs [10], in which the appearance of an unknown polymorphic form of ritonavir in drug formulations threatened the supply of the life-saving AIDS treatment Norvir, illustrates both the importance and difficulty of controlling polymorphism. In the following sections, we describe recent advances that impact industrial crystallizer control.

2.3.1 Monitoring

One of the major challenges in implementing feedback control for crystallization processes is the lack of adequate online sensors for measuring solid-state and solution properties. The United States Food and Drug Administration's (FDA) Process Analytical Technology initiative, aimed at improving pharmaceutical manufacturing practices [120], has accelerated the development and use of more advanced measurement technology. We describe several recently developed sensors for achieving better control and understanding of crystallization processes.

ATR-FTIR Spectroscopy

Attenuated total reflectance Fourier transform infrared (ATR-FTIR) spectroscopy imposes a laser beam on a sample and measures the amount of infrared light absorbed at different frequencies. The frequencies at which absorption occurs indicate which chemical species are present,

while the absorption magnitudes indicate the concentrations of these species. As demonstrated in [30] and [29], ATR-FTIR spectroscopy can be used to monitor solute concentration in a crystallization process *in situ*.

ATR-FTIR spectroscopy offers advantages over prior techniques, such as refractometry, densitometry, and conductivity measurements, for measuring solute concentration. Refractometry works only if there is a significant change in the refractive index with solute concentration and is sensitive to air bubbles. Densitometry requires sampling of the crystal slurry and filtering out the crystals to accurately measure the liquid-phase density. This sampling process involves an external loop that is sensitive to temperature fluctuations and subject to filter clogging. Conductivity measurements, which are useful only for electrolytes, require frequent re-calibration. ATR-FTIR spectroscopy overcomes these problems and can measure multiple solute concentrations. Calibration of ATR-FTIR is usually rapid [62] and thus well suited for batch processes and short production runs. In [105], linear chemometrics is applied to estimate solute concentration with high accuracy (within 0.12%). Several applications for which ATR-FTIR monitoring is useful are described in [34].

Unfortunately, ATR-FTIR spectroscopy is considerably more expensive than the alternatives. Another drawback of ATR-FTIR is the vulnerability of the IR probe's optical material to chemical attack and fouling [28].

Raman spectroscopy

Raman spectroscopy imposes a monochromatic laser beam on a sample and measures the amount of light scattered at different wavelengths. The differences in wavelength between the incident light and the scattered light is a fingerprint for the types of chemical bonds in the sample. Raman spectroscopy has been used to make quantitative polymorphic composition measurements since 1991 [25]. This technology has been applied to quantitative, *in situ* polymorphic composition monitoring in solution crystallization since 2000 [110]–[79].

Raman spectroscopy is well suited to *in situ* polymorphism monitoring for several reasons. Specifically, Raman analysis does not require sample preparation; the Raman signal can be propagated with fiber optics for remote sensing; and Raman sampling probes are less chemically sensitive than ATR-FTIR probes [28]. In addition, this technique can be used to monitor the solid and liquid phases simultaneously [32],[49].

Like ATR-FTIR, Raman-based technologies are expensive. Furthermore, calibration of the Raman signal for quantitative polymorphic composition measurements can be difficult because the signal intensity is affected by the particle size distribution. Hence Raman's utility for quantitative monitoring depends on corrections for particle-size effects [81].

Near-Infrared Spectroscopy

Near-infrared (NIR) spectroscopy is also used to quantitatively monitor polymorphic composition [78]. Like Raman, NIR is well suited for *in situ* analysis. The main drawback of NIR is that

calibration is difficult and time consuming. In some cases, however, coarse calibration is sufficient to extract the needed information [35].

Laser backscattering

Laser backscattering-based monitoring technology, such as Lasentec's FBRM probe, has proven useful for characterizing particle size and for determining saturation concentrations and metastable limits [8],[9]. This sensor measures particle *chord lengths* (see Figure S8) by moving a laser beam at high velocity through the sample and recording the *crossing times*, that is, the time durations over which light is backscattered as the laser passes over particles. The chord length of each particle traversed by the laser is calculated as the product of the laser's velocity and the crossing time of the particle. This technique allows rapid, calibration-free acquisition of thousands of chord-length measurements to robustly construct a chord length distribution (CLD). Laser backscattering technology can be applied *in situ* under high solids concentrations.

Because laser-backscattering provides a measurement of only chord length, this technique cannot be used to measure particle shape directly. Also, inferring the PSD from the CLD involves the solution of an ill-posed inversion problem. Although methods for solving this inversion problem are developed in [63],[116], these methods depend on assumptions regarding particle shape. CLD thus provides only a qualitative indication of the underlying PSD characteristics.

Video Microscopy

Video microscopy can be used to characterize both crystal size and shape. Furthermore, for chemical systems in which the polymorphs exhibit different shapes, such as glycine in water, video microscopy can be used to monitor polymorphic composition [18]. Obtaining all three of these measurements using a single probe reduces cost and simplifies the experimental setup. Video microscopy is also appealing because interpretation of image data is intuitive.

Several challenges associated with video microscopy have limited its application for particle size and shape analysis. Most commercial video-microscopy-based analyzers require sampling to obtain images for automatic image analysis. Although *in situ* probes are available, their utility has mainly been limited to qualitative monitoring because the nature of the images, which contain blurred, out-of-focus, and overlapping particles, precludes the successful application of image analysis to automatically quantify particle size and shape. Furthermore, crystal-size measurements obtained from images depend on the orientation of the crystals with respect to the camera, although the effect is less pronounced than for laser backscattering-based measurements (see Figure S8).

2.3.2 Manipulated variables

In advanced control implementations, the quality variables of interest (crystal size, shape, form, and purity) are indirectly controlled using manipulated variables that affect the supersaturation level in the crystallizer. For example, cooling crystallizers manipulate the crystallizer temper-

ature to change the saturation concentration of the crystallizing species. Anti-solvent crystallizers change the saturation concentration by manipulating the solvent composition.

Controlling the supersaturation level provides only limited control over the resulting crystal shape distribution, size distribution, and polymorphic form. Progress is being made in this area as well, however. Several research groups are investigating additives that bind to selected crystal faces to inhibit growth of the faces, thereby promoting a desired crystal shape or polymorphic form [111],[13]. These additives, which are similar to the target-crystallizing species, can be incorporated onto a growing crystal face. The remaining exposed portion of the additive consists of chemical groups that hinder further growth on that face. Additives are also used as nucleation promoters or inhibitors to obtain a desired polymorphic form [112]. Templates, such as single-crystal substrates, are also being investigated as a means for manipulating nucleation events to obtain desired polymorphs [75].

Additives are used in industry to help understand the effect of impurities on crystal growth and polymorphism [13]. As yet, additives and templates are not widely used as manipulated variables in industrial crystallizer control schemes. To use these methods as manipulated variables for advanced control, models are needed to describe the effect of the manipulated variables on crystal size, shape, and form.

2.3.3 Modeling

Process Modeling

Crystallizers have highly nonlinear, complex dynamics including multiple steady states, open-loop instability, and long time delays; hence low-order, linear models are often inadequate for control purposes. Furthermore, nonlinear black box models, such as neural networks, are also inadequate in many cases because batch crystallizers have such a large operating region. Online, optimal control of dynamic crystallization is thus enhanced by the ability to efficiently simulate the underlying physical model, which is a system of partial integro-differential equations that couples mass and energy balances with a population balance describing the evolution of the crystal population's PSD.

The evolution of the PSD often involves sharp, moving fronts that are difficult to simulate efficiently. Several software packages are now available for simulating crystallization processes. The commercial software package PARSIVAL [118] is designed to handle the entire class of crystallizer configurations and crystallization phenomena. AspenTech's process simulation software has specific tools for crystallization simulation and troubleshooting. The crystallizer simulator of the GPROMS package can be interfaced with computational fluid dynamics (CFD) software, such as FLUENT or STAR-CD. DYNOCHEM solves the population balance equations by applying the commonly used method of moments, a model reduction technique. Researchers have developed techniques that extend the applicability of the method of moments to systems involving length-dependent crystal growth [70],[76], and these developments are being incorporated directly into CFD software packages [97].

Shape Modeling

Models and methods for predicting crystal shape based solely on knowledge of the internal crystal structure are available in software packages such as CERIUS^2 and HABIT [22]. These methods provide accurate shape predictions for vapor-grown crystals but, for solution-grown crystals, do not take into account the effects of supersaturation, temperature, solvent, and additives or impurities. Current shape-modeling research is focused on accounting for these effects [114]–[93].

Polymorphism Modeling

The ability to predict the crystal structure for a given molecule in a given environment would represent a major advance in drug development. Although significant progress has been made in making these predictions [87],[24], this problem is far from being solved. Most approaches seek to find the crystal structure that corresponds to the global minimum in lattice energy, that is, the most thermodynamically stable form at zero Kelvin. These approaches neglect entropic contributions arising at higher temperatures as well as kinetic effects due to the experimental crystallization conditions. Current polymorphism modeling methods thus cannot reliably predict the polymorphs that are observed experimentally.

2.3.4 Control

The developments described above impact the way crystallizer control is approached in industry. In particular, the development of better measurement technology enables the application of simple but effective crystallizer control strategies. Most of these strategies focus on feedback control of supersaturation using concentration (by means of ATR-FTIR) and temperature measurements to follow a predefined supersaturation trajectory [64]–[38]. This approach is attractive because it can be implemented without characterizing the nucleation and growth kinetics, using only the saturation concentration and metastable zone width data. Furthermore, this approach results in a temperature profile that can be used in large-scale crystallizers that do not have concentration-measurement capabilities.

Laser backscattering measurements are used to control the number of particles in the system [27]. This control strategy alternates between cooling and heating stages, allowing the heating stage to continue until the particle count number measured by FBRM returns to its original value upon seeding, indicating that all fine particles generated by secondary nucleation have been dissolved. In [84], online crystal shape measurements obtained by means of optical microscopy and automated image processing are used to manipulate impurity concentration and thereby control crystal habit. The control schemes employed in these experimental studies are basic (usually PID or on-off control), illustrating that, given adequate measurement technology, simple control schemes can often provide an adequate level of control capability.

More sophisticated control methods have also been demonstrated for batch crystallizers. Experimental results obtained in [74]–[117] demonstrate product improvement by using predictive, first-principles models to determine open-loop, optimal cooling and seeding policies. Closed-loop, optimal control of batch crystallizers is demonstrated using simulations in [92],[100].

For continuous crystallizers, various model-based, feedback controllers have been suggested. In [109], an H_∞ controller based on a linearized distributed parameter model is shown to successfully stabilize oscillations in a simulated, continuous crystallizer using measurements of the overall crystal mass, with the flow rate of fines (small crystals) to the dissolution unit as the manipulated variable. In [100], a hybrid controller combining model predictive control with a bounded controller is used to ensure closed-loop stability for continuous crystallizers.

2.4 The future of crystallization

As the chemical, pharmaceutical, electronics, and food industries continue to develop new products, crystallization will enjoy increasingly wide application as a means to separate and purify chemical species and create solids with desirable properties. Sensor technology for crystallizers will continue to improve, especially given the current emphasis by the FDA's Process Analytical Technology initiative. Industries that have used batch crystallizers to produce low-volume, high-value-added chemicals might choose to move to continuous crystallizers to reduce operating costs and enable more flexible and compact process design. Shape modeling and molecular modeling tools offer tremendous potential for enabling robust process design and control, and these tools can be expected to advance rapidly given the amount of interest and research in this area. These developments will impact the way in which crystallization processes are designed and will enable more effective control of size distribution, shape, and polymorphic form, leading to the creation of crystalline solids that are useful for a wide variety of applications.

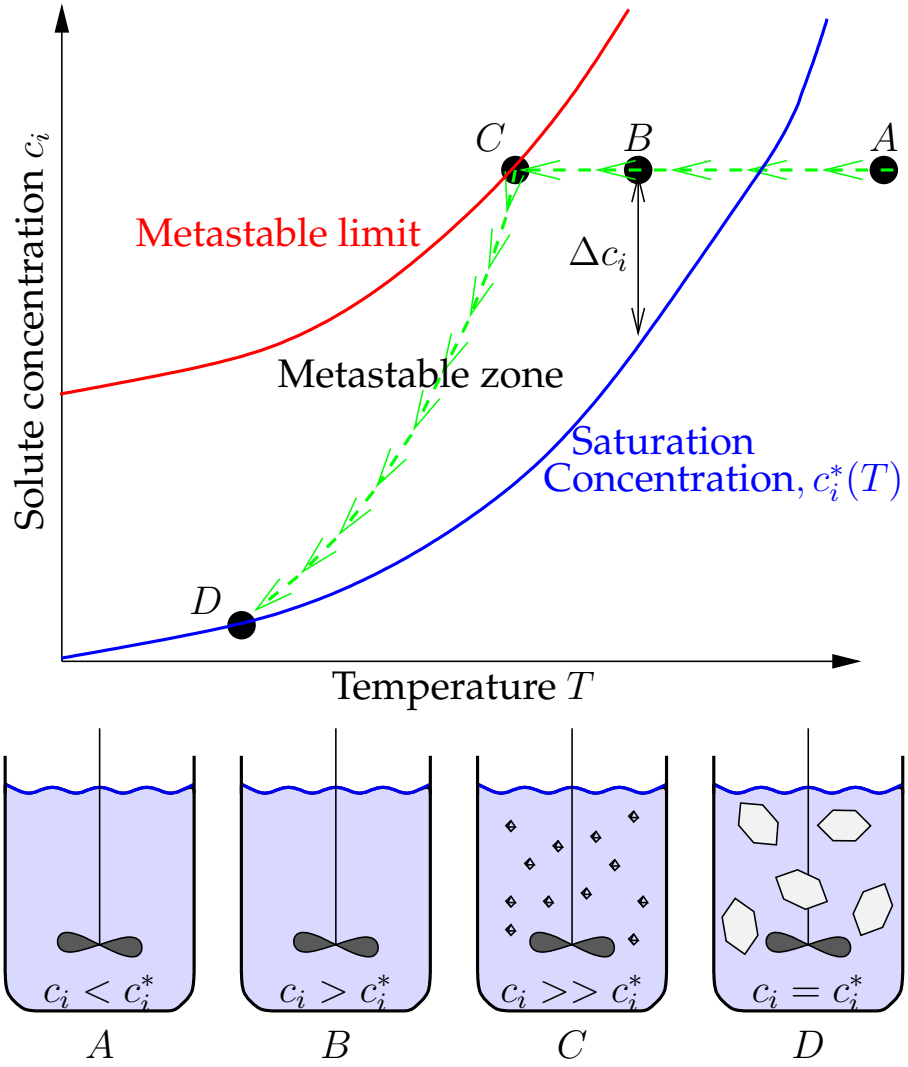


Figure S1: Depiction of a cooling, solution crystallization process. The process begins at point A, at which the solution is undersaturated with respect to species i ($c_i < c_i^*$). The process is cooled to point B, at which the solution is supersaturated ($c_i > c_i^*$). No crystals form at point B, however, because the activation energy for nucleation is too high. As the process cools further, the supersaturation level increases and the activation energy for nucleation decreases. At the metastable limit (point C), spontaneous nucleation occurs, followed by crystal growth. The solute concentration decreases as solute molecules are transferred from the liquid phase to the growing crystals until equilibrium is reached at point D, at which $c_i = c_i^*$.

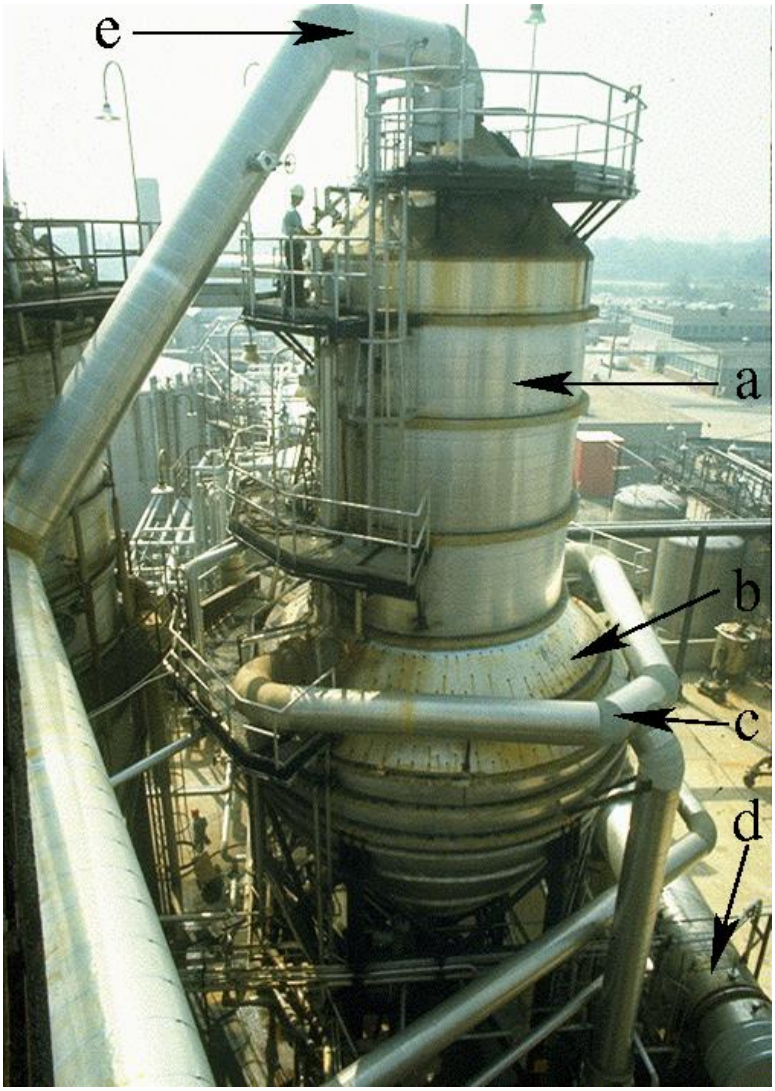


Figure S2: Production-scale draft tube crystallizer. This crystallizer is used to produce hundreds of tons per day of ammonium sulfate, commonly used as fertilizer or as a precursor to other ammonium compounds. The crystallizer body (a) widens at the lower section (b) to accommodate the settling region, in which small crystals called fines are separated from the larger crystals by gravitational settling. The slurry of saturated liquid and fines in the settling region is continuously withdrawn (c), combined with product feed, and passed through a heater (d) that dissolves the fines and heats the resulting solution prior to returning the solution to the crystallizer. The heat generated by crystallization is removed as the solvent evaporates and exits through the top of the crystallizer (e), to be condensed and returned to the process. Larger crystals are removed continuously from the bottom of the crystallizer. Image courtesy of Swenson Technology, Inc.

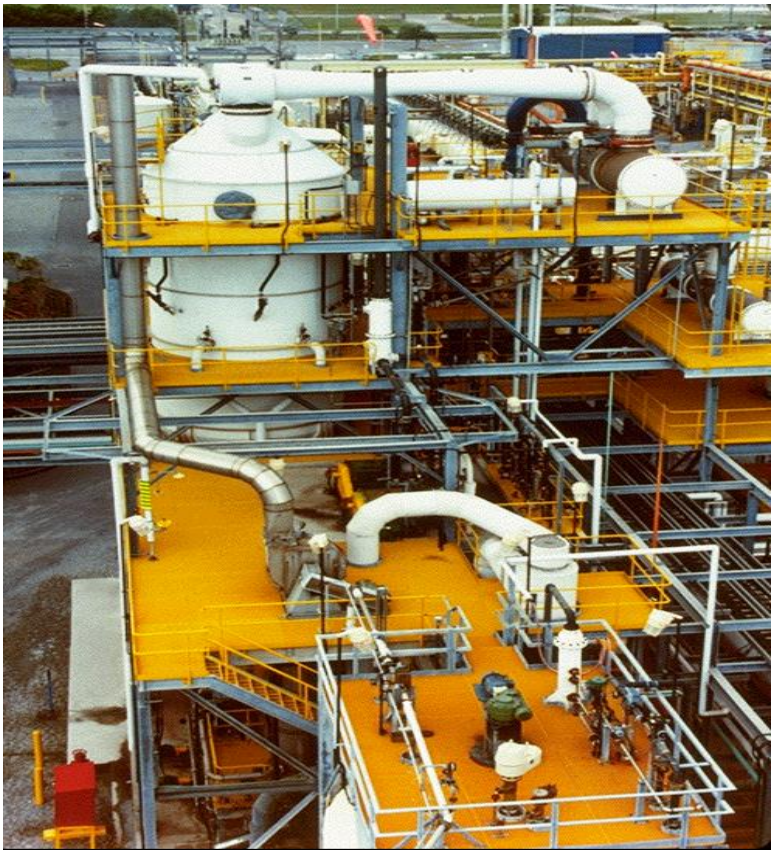


Figure S3: Production-scale draft tube baffle crystallizer. This crystallizer is used to produce hundreds of tons per day of sodium chlorate, which is commonly used in herbicides. Image courtesy of Swenson Technology, Inc.

Figure S4: Small crystallizer used for high potency drug manufacturing. The portal (a) provides access to the crystallizer internals. The crystallizer widens at the lower section (b) to accommodate the crystallizer jacket, to which coolant (e) and heating fluid (h) lines are connected. Mixing is achieved using an impeller driven from below (c). The process feed enters from above (f) and exits below (d). The temperature sensor is inserted from above (g). Image courtesy of Ferro Pfanstiehl Laboratories, Inc.

Figure S5: Upper section (top image), lower section (center image), and internals of batch crystallizer, showing the impeller and temperature sensor. This crystallizer is used for contract pharmaceutical and specialty chemical manufacturing. Images courtesy of Avecia, Ltd.

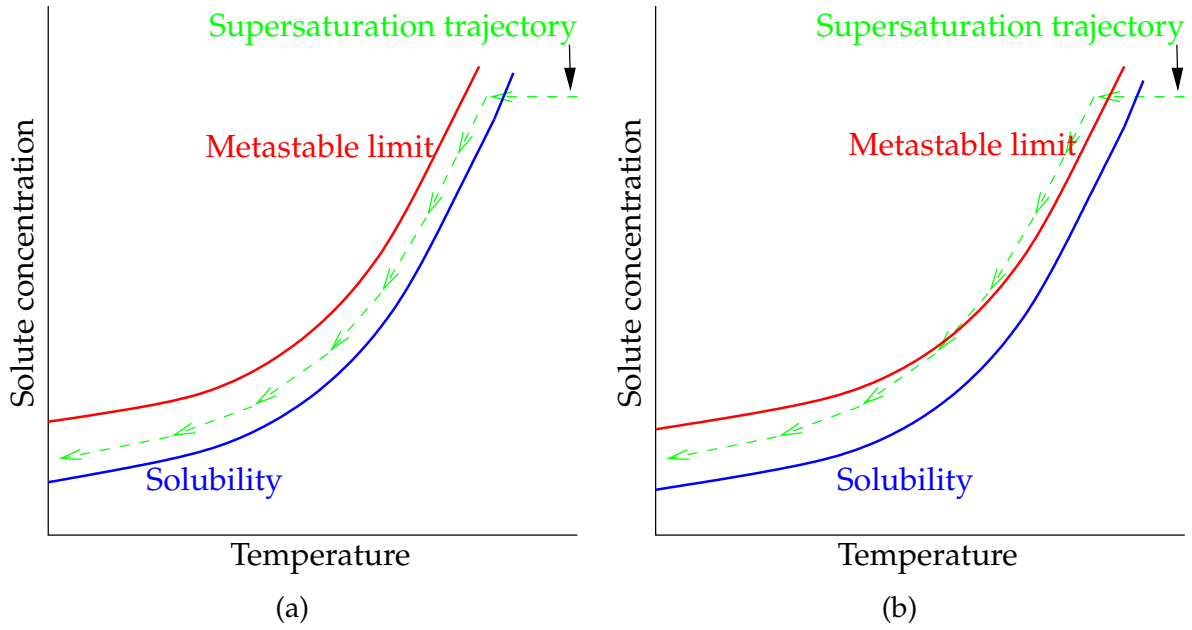


Figure S6: Batch cooling crystallization. In this illustration, the process is cooled until it becomes supersaturated and crystallization can occur. As the solute species deposit onto the forming crystals, the solute concentration decreases. Supersaturation is therefore maintained by further cooling. As shown in (a), a well-controlled crystallization process operates in the metastable zone between the saturation concentration and metastable limit, balancing the nucleation and growth rates to achieve the desired crystal size distribution. As depicted in (b), disturbances such as impurities can shift the metastable zone, resulting in undesired nucleation that substantially degrades the resulting particle size distribution.

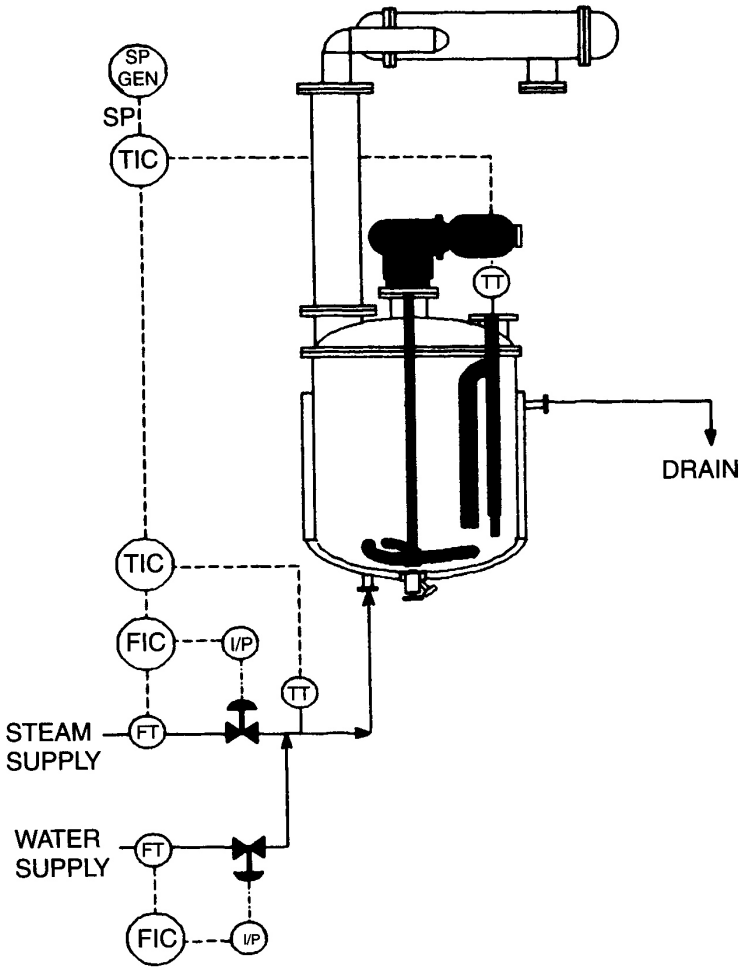


Figure S7: General purpose glass-lined batch reactor/crystallizer commonly used for producing fine chemicals and pharmaceuticals. The primary controller compares the measured solution temperature with a prespecified setpoint trajectory to calculate the temperature setpoint for the cooling jacket. The secondary controller manipulates the flow of steam to the jacket to reach the jacket temperature setpoint. Reprinted from [90], with permission from Elsevier.

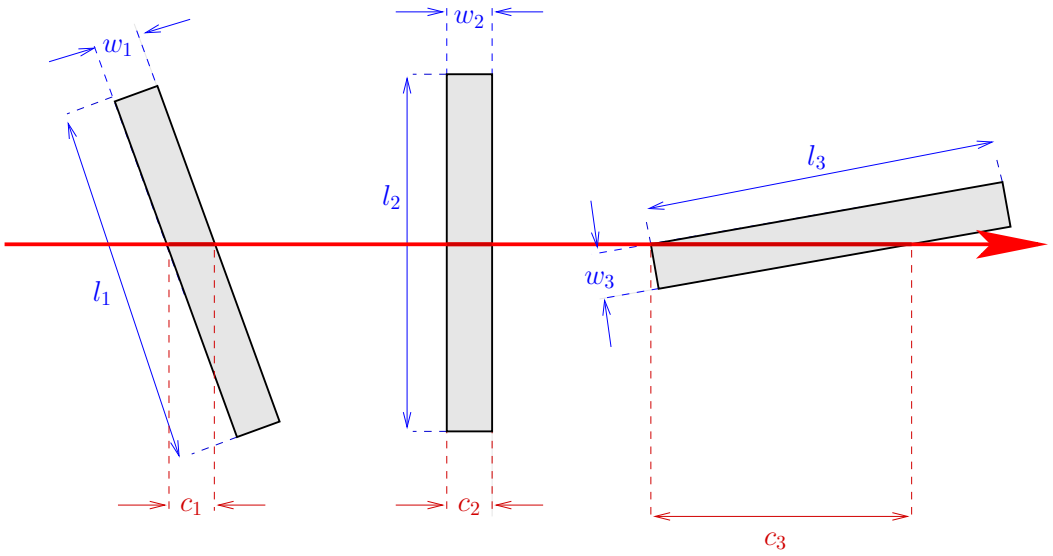


Figure S8: Comparison of crystal-size measurements obtained using laser backscattering versus those obtained using vision. Laser backscattering provides chord lengths (c_1, c_2, c_3) while vision-based measurement provides projected lengths (l_1, l_2, l_3) and projected widths (w_1, w_2, w_3). The chord-length measurement for each particle depends on its orientation with respect to the laser path (depicted above using the red arrow), while the projected length and width measurements are independent of in-plane orientation. Size measurements from both techniques are affected by particle orientation in depth.

Chapter 3

Particle size distribution model

3.1 Model formulation

3.1.1 Population balance

3.1.2 Mass balance

3.1.3 Energy balance

3.2 Model solution

3.2.1 Method of moments

3.2.2 Orthogonal collocation

The orthogonal collocation method, as explained in [108], consists of approximating the model solution at each time step as a nth order polynomial such that the spatial derivatives in equations **??** and **??** can be approximated as linear combinations of the model solution values at n collocation locations along the spatial domain, i.e.

$$\left.\frac{dc}{dz}\right|_{z_i} = \sum_{j=1}^{n_c} A_{ij} c_j \tag{3.1}$$

$$\left.\frac{d^2c}{dz^2}\right|_{z_i} = \sum_{j=1}^{n_c} B_{ij} c_j \tag{3.2}$$

$$\tag{3.3}$$

in which $c_j = c(z_j, t)$ and n_c is the number of collocation points. This approximation reduces the PDE's in equations **??** and **??** and the boundary conditions in **??** and **??** to the following set of

DAE's:

$$0 = c_{il} + \frac{u}{\epsilon}\sum_{j=1}^{n_c} A_{ij}c_{jl} + Fk_f(Kc_{il} - q_{il}) + D_L\sum_{j=1}^{n_c} B_{ij}c_{jl} \tag{3.4}$$

$$0 = q_{il} - k_f(q^*_{il} - q_{il}) \tag{3.5}$$

$$0 = D_L\sum_{j=1}^{n_c} A_{1j}c_{jl} + \frac{u}{\epsilon}(c_{F_l} - c_{1l}) \tag{3.6}$$

$$0 = \sum_{j=1}^{n_c} A_{n_cj}c_{jl} \tag{3.7}$$

in which $i = 2, 3, \ldots, n_c$ and $l = 1, 2, \ldots, n_s$ (n_s is the number of species) for equations 3.4 and 3.5 (the first subscript on the quantities c, c, q, and q refers to the index of the collocation point, or collocation location and the second subscript refers to the species). The COLLOC function is used to obtain the collocation locations ($z_j, j = 1, 2, \ldots, n_c$) and the derivative matrices ($\mathbf{A}$ and $\mathbf{B}$). The DASRT function, based on Petzold's DASSL software [85], is used to integrate the DAE's.

If the model solution involves steep concentration gradients, application of orthogonal collocation on the entire integration domain requires the use of a very high order polynomial to approximate the solution, which usually leads to a highly oscillatory solution. Therefore, it may be necessary to divide up the integration domain into finite elements and apply collocation on each element while imposing certain constraints on the solution values and derivatives at the elements' boundaries, a procedure referred to as either "global spline collocation" [108] or "orthogonal collocation on finite elements" [36]. This method has been applied to the chromatography model both by using stationary elements (a procedure termed "orthogonal collocation on fixed elements," or OCFE) [67] and moving elements [121]. The latter approach requires the integration to be interrupted on a regular basis to examine the current model solution and determine optimal locations for the elements' boundaries based on gradient information. These interruptions cause significant time delays. The prior approach can also be slow since every element must be able to handle the steep gradients, meaning that either each element must have a large number of collocation points or many elements must be used.

The approach used in this study is similar to the moving element approach mentioned above with the difference that the elements' boundary locations evolve in time according to analytical expressions. This precludes the necessity of continually interrupting the integration process, thereby resulting in much faster simulation. A similar formulation is used in [39] to simulate a fixed bed catalytic reactor.

Implementing this approach requires a change from the fixed coordinate z, which denotes axial position along the entire column, to the moving coordinate z, which denotes axial position within each finite element:

$$z(z,t) = \frac{z - Z_k(t)}{Z_{k+1}(t) - Z_k(t)} = \frac{z - Z_k(t)}{\Delta_k(t)} \tag{3.8}$$

in which Z_{k+1} and Z_k are, respectively, the upper and lower bounds of the kth moving element. Hence, the difference between them, Δ_k, is the width of the kth element. Each boundary location

Z_k evolves according to a pre-specified function g_k (see Section **??**):

$$\frac{dZ_k}{dt} = g_k(t) \tag{3.9}$$

The mobile and stationary phase concentrations are accordingly redefined as

$$\begin{aligned} c(z,t) &= c(z,t) && (3.10)\\ q(z,t) &= q(z,t) && (3.11)\\ q^*(c) &= q^*(c) && (3.12) \end{aligned}$$

such that

$$\begin{aligned} \frac{\partial c}{\partial t} &= \frac{\partial c}{\partial z}\frac{\partial z}{\partial t} + \frac{\partial c}{\partial t} && (3.13)\\ \frac{\partial q}{\partial t} &= \frac{\partial q}{\partial z}\frac{\partial z}{\partial t} + \frac{\partial q}{\partial t} && (3.14)\\ \frac{\partial c}{\partial z} &= \frac{\partial c}{\partial z}\frac{\partial z}{\partial z} && (3.15) \end{aligned}$$

It can be shown that

$$\begin{aligned} \frac{\partial z}{\partial t} &= -\frac{z(g_{k+1}-g_k)+g_k}{\Delta_k} && (3.16)\\ \frac{\partial z}{\partial z} &= \frac{1}{\Delta_k} && (3.17) \end{aligned}$$

Equations 3.13-3.17 can be substituted into equations **??** and **??** to obtain the model in the transformed coordinate system:

$$\frac{\partial c_i}{\partial t} + \frac{1}{\Delta_k}\left(\frac{u}{\epsilon} - z(g_{k+1}-g_k) - g_k\right)\frac{\partial c_i}{\partial z} - \frac{D_L}{\Delta_k^2}\frac{\partial^2 c_i}{\partial z^2} + \frac{1-\epsilon}{\epsilon}k_f\left(q_i^* - q_i\right) = 0 \tag{3.18}$$

$$\frac{\partial q_i}{\partial t} - \left(\frac{z(g_{k+1}-g_k)+g_k}{\Delta_k}\right)\frac{\partial q_i}{\partial z} - k_f\left(q_i^* - q_i\right) = 0; \tag{3.19}$$

The initial conditions are

$$\begin{aligned} c(z,0) &= 0 && (3.20)\\ q(z,0) &= 0 && (3.21) \end{aligned}$$

$$Z_k(0) = \begin{cases} 0 & k = 1,2,\ldots,n_b-1 \\ L & k = n_b \end{cases} \tag{3.22}$$

The boundary conditions are obtained by transforming the original boundary conditions at the column inlet and outlet and imposing continuity and smoothness at the boundaries between

elements:

$$\frac{D_L}{Z_2 - Z_1} \left.\frac{\partial c_i}{\partial z}\right|_{z=0,t} = -\frac{u}{\epsilon}\left[c_i(0^-,t) - c_i^1(0^+,t)\right] \tag{3.23}$$

$$\left.\frac{\partial c_i^{n_e}}{\partial z}\right|_{z=1,t} = 0 \tag{3.24}$$

$$c_i^k(1,t) = c_i^{k+1}(0,t), \quad k = 1,2,\ldots n_e - 1 \tag{3.25}$$

$$\frac{1}{\Delta_k}\left.\frac{\partial c_i^k}{\partial z}\right|_{z=1,t} = \frac{1}{\Delta_{k+1}}\left.\frac{\partial c_i^{k+1}}{\partial z}\right|_{z=0,t}, \quad k = 1,2,\ldots n_e - 1 \tag{3.26}$$

in which $c_i(0^-,t)$ is defined as

$$c_i(0^-,t) = \begin{cases} c_{F_i} & 0 < t \le t_F \\ 0 & t > t_F \end{cases} \tag{3.27}$$

The introduction of a $\frac{\partial q}{\partial z}$ term in the transformed coordinate system requires the specification of a single boundary condition for q on each element. These are obtained by imposing continuity at the boundaries between elements:

$$q_i^k(1,t) = q_i^{k+1}(0,t), \quad \text{k = 1, 2, } \ldots n_e - 1 \tag{3.28}$$

This leaves one more boundary condition to be specified at the column outlet. However, since the value of q_i is not known at the outlet, it will have to suffice to use the following condition:

$$\left.\frac{\partial q_i^{n_e}}{\partial z}\right|_{z=1,t} = 0 \tag{3.29}$$

Application of the collocation procedure to these equations results in the following set of DAE's:

$$0 = \frac{dc_{il}^k}{dt} + \frac{1}{\Delta_k}\left(\frac{u}{\epsilon} - z_i^k(g_{k+1} - g_k) - g_k\right)\sum_{j=1}^{n_c^k} A_{ij}^k c_{jl}^k - \frac{D_L}{\Delta_k^2}\sum_{j=1}^{n_c^k} B_{ij}^k c_{jl}^k + \frac{1-\epsilon}{\epsilon} k_f\left(q_{il}^{*k} - q_{il}^k\right), \quad i = 2,3,\ldots,n_c^k - 1; \quad k = 1,2,\ldots,n_e \tag{3.30}$$

$$0 = \frac{dq_{il}}{dt} - \left(\frac{z_i^k(g_{k+1} - g_k) + g_k}{\Delta_k}\right)\sum_{j=1}^{n_c^k} A_{ij}^k q_{jl}^k - k_f\left(q_{il}^{*k} - q_{il}^k\right), \quad i = 1,2,\ldots,n_c^k - 1; \quad k = 1,2,\ldots,n_e \tag{3.31}$$

$$0 = \frac{D_L}{\Delta_1}\sum_{j=1}^{n_c^1} A_{1j}^1 c_{jl}^1 + \frac{u}{\epsilon}\left(c_l(0^-,t) - c_{1l}^1\right) \tag{3.32}$$

$$0 = \sum_{j=1}^{n_c^{n_e}} A_{ij}^{n_e} c_{jl}^{n_e} \tag{3.33}$$

$$0 = c_{n_c l}^k - c_{1l}^{k+1}, \quad k = 1, 2, \ldots, n_e - 1 \tag{3.34}$$

$$0 = \frac{1}{\Delta_k}\sum_{j=1}^{n_c^k} A_{n_c j}^k c_{jl}^k - \frac{1}{\Delta_{k+1}}\sum_{j=1}^{n_c^{k+1}} A_{1j}^{k+1} c_{jl}^{k+1}, \quad k = 1, 2, \ldots, n_e - 1 \tag{3.35}$$

$$0 = q_{n_c l}^k - q_{1l}^{k+1}, \quad k = 1, 2, \ldots, n_e - 1 \tag{3.36}$$

$$0 = \sum_{j=1}^{n_c^{n_e}} A_{ij}^{n_e} q_{jl}^{n_e} \tag{3.37}$$

in which the superscripts and subscripts are interpreted as follows: c_{il}^k refers to the concentration of species l at the ith collocation location of the kth element. Similarly, the superscript on the collocation matrices refers to the element index.

3.2.3 Parameter Estimation

This section explains the general approach used in analyzing the CIAT parameter estimation problem. The approach and derivations given in this section are summarized from *Chemical Reactor Analysis and Design Fundamentals* [**?**] and are applicable to the case of a nonlinear differential equation model of the form

$$\frac{d\mathbf{x}}{dt} = \mathbf{f}(\mathbf{x}, \theta) \tag{3.38}$$

$$\mathbf{x}(0) = \mathbf{g}(\mathbf{x}_0, \theta) \tag{3.39}$$

$$\mathbf{y} = \mathbf{h}(\mathbf{x}(\theta), \theta)) \tag{3.40}$$

in which $\mathbf{x}$ represents the vector of states, $\mathbf{y}$ represents the vector of measurable variables, and θ represents the vector of parameters being estimated.

The goal of parameter estimation is two-fold: (1) To determine model parameter values that allow the model to accurately predict experimental behavior, and (2) to determine the level of confidence in those values. To achieve the first goal, we seek to minimize an objective function of the form

$$\Phi(\theta) = \sum_i \mathbf{e}_i^T \mathbf{W} \mathbf{e}_i \tag{3.41}$$

$$\mathbf{e}_i = \mathbf{y}_i - \mathbf{h}(\mathbf{x}(t_i, \theta), \theta) \tag{3.42}$$

in which $\mathbf{e}_i$ represents the vector of differences between the data ($\mathbf{y}_i$) and the model predictions at time t_i, and $\mathbf{W}$ is a symmetric, positive-definite matrix used to weight the data-model mismatch. The optimal value of θ, or the value of θ for which Equation 3.41 is minimized, will be denoted θ.

To achieve the second goal, we need to identify the range of values over which θ can vary without causing a statistically significant change in objective function. That is, we are looking for values of θ for which

$$|\Phi(\theta) - \Phi(\theta)| \leq s^2 n_p F(n_p, n_d - n_p, \alpha) \tag{3.43}$$

in which s^2 is the sample variance, n_p is the number of parameters being estimated, n_d is the number of data points, α is the level of confidence (in this report, we use 95%), and F represents the F-distribution. To convert the left hand side of Equation 3.43 into a more analytical form, we perform a second-order, multivariable Taylor series expansion of the objective function at the optimal solution, θ. Assuming the gradient is zero at the optimal solution, we can rearrange the Taylor expansion to give

$$\Phi(\theta) - \Phi(\theta) \approx \frac{1}{2} (\theta - \theta)^T \mathbf{H}\big|_{\theta=\theta} (\theta - \theta) \tag{3.44}$$

in which $\mathbf{H}$, the Hessian, is the second derivative of the objective function with respect to the model parameters (see Section **??**). We combine Equations 3.43 and 3.44 to obtain an equation for calculating approximate confidence intervals:

$$(\theta - \theta)^T \mathbf{H}\big|_{\theta=\theta} (\theta - \theta) \leq 2 s^2 n_p F(n_p, n_d - n_p, \alpha) \tag{3.45}$$

The most conservative confidence interval for each individual parameter is obtained by assuming that the other parameters are at their optimal values, in which case Equation 3.45 can be rearranged to give

$$\theta_i = \theta_i \pm \sqrt{2 s^2 n_p F(n_p, n_d - n_p, \alpha) V_{ii}} \tag{3.46}$$

in which V_{ii} is the ith diagonal element of the inverse of the Hessian and the sample variance s^2 is calculated using

$$s^2 = \frac{\Phi(\theta)}{n_d - n_p} \tag{3.47}$$

The Hessian is a square, symmetric matrix with elements given by

$$H_{kj} = \frac{\partial^2 \Phi}{\partial \theta_k \partial \theta_j} \tag{3.48}$$

The calculation of the Hessian is facilitated by using the Gauss-Newton approximation, which assumes the data-model mismatch is small and randomly distributed about zero. Using this approximation, the Hessian can be calculated using

$$\mathbf{H} = 2 \sum_i \left(\mathbf{S}_i^T \frac{\partial \mathbf{h}_i^T}{\partial \mathbf{x}_i} + \frac{\partial \mathbf{h}_i^T}{\partial \theta} \right) \mathbf{W} \left(\frac{\partial \mathbf{h}_i}{\partial \theta^T} + \frac{\partial \mathbf{h}_i}{\partial \mathbf{x}_i^T} \mathbf{S}_i \right) \tag{3.49}$$

where, again, the subscript i represents the value of the variable at time t_i and $\mathbf{S}$ is the time-varying matrix of sensitivities given by

$$S_{jk}(t_i) = \frac{\partial x_j(t_i, \theta)}{\partial \theta_k} \tag{3.50}$$

The Hessian is useful for analyzing confidence intervals not only because of its presence in Equation 3.46, but also because its eigenvalues and eigenvectors give us information about the optimization surface. The connection between the eigenvalues, the eigenvectors, and the optimization surface (i.e. objective function surface) is demonstrated in Figure 3.2.3, which shows a plot of Equation (3.44) for a two-parameter case ($\delta\theta$ has been substituted for $\theta - \theta$). The ellipse in Figure 3.2.3 is a contour of the objective function at $b = 2(\Phi(\theta) - \Phi(\theta))$. The λ's and the vectors $\mathbf{v}$ and $\mathbf{w}$ represent the eigenvalues and eigenvectors of the Hessian, respectively. This figure demonstrates that, for a given change in objective function, we can move a long distance in the direction of $\mathbf{v}$ but only a small distance in the direction of $\mathbf{w}$. The two distances are inversely related to the corresponding eigenvalues. In other words, the eigenvalues and eigenvectors of the Hessian tell us how much, and in which direction, we can move the parameters before causing a given change in objective function and therefore give us a quantitative and quick way to examine both correlation between parameters and confidence intervals. This information is useful for model reduction, as explained in the following section.

Assume that, for a given case, an eigenvector/eigenvalue analysis (see Section **??**) indicates that the data do not contain sufficient information to place tight confidence intervals on the absolute values of the parameters θ_1 and θ_2. However, the data seem to contain information regarding a correlation between these two parameters. In this case, the model can possibly be reduced such that a tight confidence interval can be found on the value of some *combination* of these parameters rather than their individual, absolute values. As explained by Bard [**?**], this is accomplished in the following manner.

Let ϵ be the difference between objective values that we consider statistically insignificant. The values of θ that we will accept are those for which

$$|\Phi(\theta) - \Phi(\theta)| \leq \epsilon \tag{3.51}$$

in which θ is the vector of optimal parameter values. We have already shown that

$$\Phi(\theta) - \Phi(\theta) \approx \frac{1}{2} (\delta\theta)^T \mathbf{H}\big|_{\theta=\theta} (\delta\theta) \tag{3.52}$$

in which $\delta\theta$ is defined as

$$\delta\theta = \theta - \theta \tag{3.53}$$

In $\delta\theta_1$-$\delta\theta_2$ space, Equation (3.52) gives an elliptical "indifference region" such as that shown in Figure 3.2.3. We can visit any point in the indifference region using the equation

$$\begin{bmatrix} \delta\theta_1 \\ \delta\theta_2 \end{bmatrix} = \alpha \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} + \beta \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} \tag{3.54}$$

in which $\mathbf{v}$ and $\mathbf{w}$ are the orthogonal, normalized eigenvectors of the Hessian, pointing in the direction of the ellipse's major and minor axes, respectively. Multiplying both sides by $\mathbf{w}^T$ gives

$$w_1\delta\theta_1 + w_2\delta\theta_2 = \beta \tag{3.55}$$

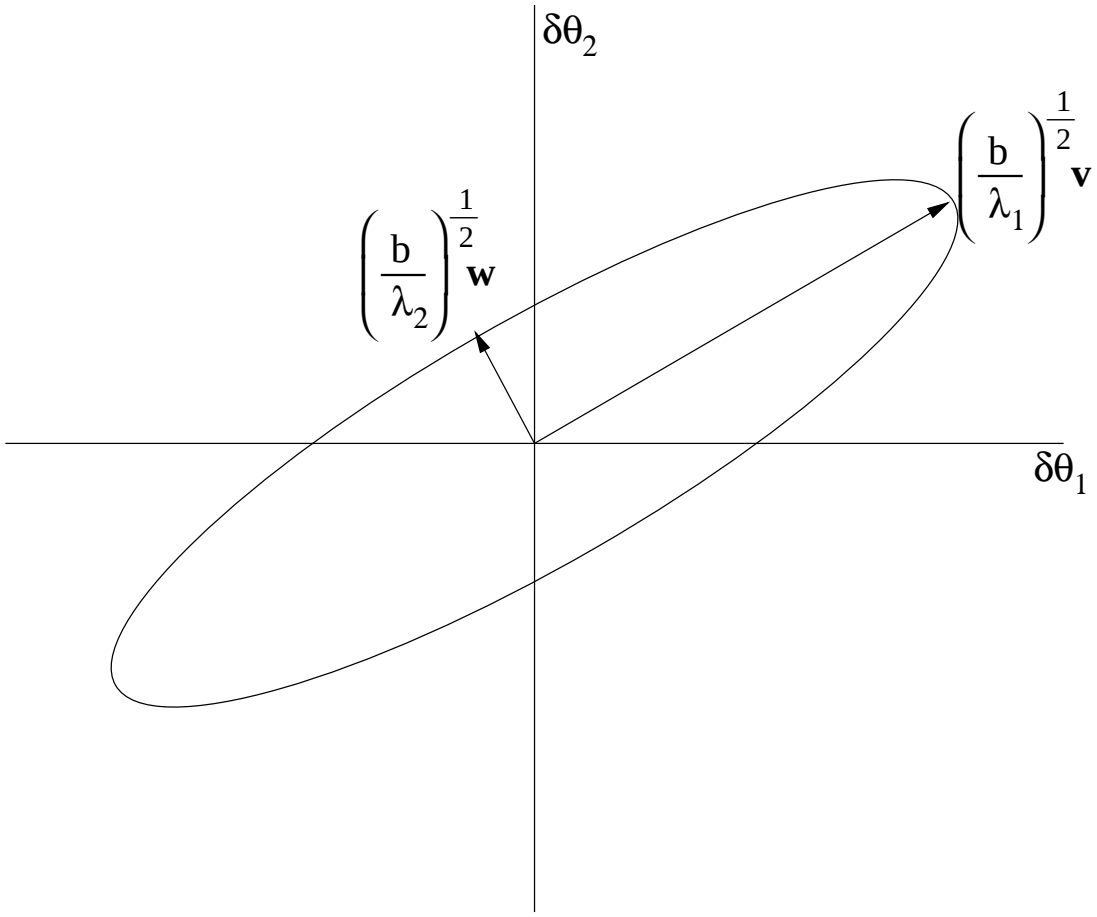


Figure S1: Geometry of the quadratic function $(\delta\theta)^T \mathbf{H}\big|_{\theta=\theta} (\delta\theta) = b$ with $\mathbf{Hv} = \lambda_1 \mathbf{v}$ and $\mathbf{Hw} = \lambda_2 \mathbf{w}$.

while multiplying both sides by $\mathbf{v}^T$ gives

$$v_1 \delta\theta_1 + v_2 \delta\theta_2 = \alpha \tag{3.56}$$

The range on β is very limited, which indicates that the relationship $(w_1 \delta\theta_1 + w_2 \delta\theta_2)$ is well-determined. On the other hand, the range on α is large, indicating that $(v_1 \delta\theta_1 + v_2 \delta\theta_2)$ is poorly determined.

To obtain a confidence interval for β, we must assign α some arbitrary value, solve for θ_1 and θ_2 in terms of β, rewrite the model equations in terms of β, and proceed with the analysis as before. The confidence interval for β will depend slightly on the value assigned to α. We obtain the most conservative confidence interval by assigning α to be zero.

Chapter 4

Experimental

4.1 Crystallizer

The experimental setup for the crystallization experiments is depicted in Figure S1. The crystallizer is a 500 mL, flat-bottomed, jacketed, glass vessel (Wilmad-LabGlass). Mixing is achieved in the crystallizer using a 3.8 cm marine-type stainless steel impeller driven by a motor controller with a speed range of 0 to 1250 revolutions per minute. A stainless steel draft tube is used to enhance mixing. The crystallizer temperature is controlled using automatic feedback control.

4.2 Measurements

4.2.1 Temperature

4.2.2 Concentration

4.2.3 Video images

4.2.4 XRPD

4.2.5 Raman

The imaging system used in this study was developed by researchers at GlaxoSmithKline and consists of a monochrome CCD video camera synchronized with a xenon strobe light and connected to a local computer via a frame grabber. Images are acquired at a rate of 30 frames per second. The camera gives images of 480 x 640 pixels and is fitted with a lens of x2 magnification providing a 280 μm depth of field and 2.48 x 1.87 mm field of view. The camera and strobe are placed to the side of the crystallizer roughly forty-five degrees apart and the images are taken through an "optical flat" attached to the side of the vessel to minimize distortion due to the curved surface of the vessel.

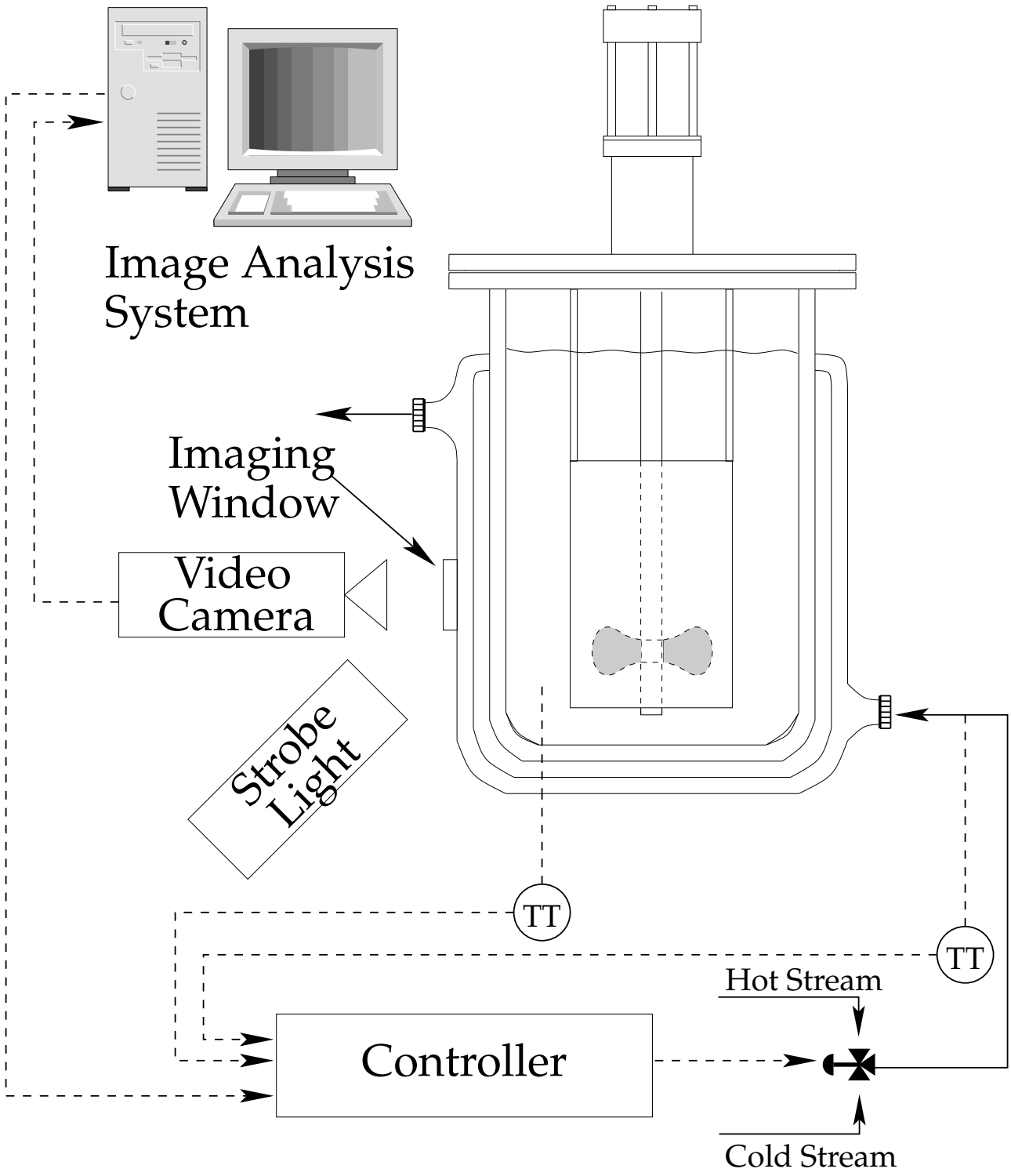


Figure S1: Experimental setup for obtaining in situ crystallization images.

4.3 Data Acquisition and Control

4.3.1 Hardware

4.3.2 Software

4.3.3 Temperature control

4.4 Chemical Systems

4.4.1 Industrial pharmaceutical

4.4.2 Glycine

Chapter 5

Two-dimensional Object Recognition for High-Aspect-Ratio Particles

Suspension crystallization processes often result in crystals having a high aspect ratio, a shape commonly described as needle-like, rod-like, or acicular. High-aspect-ratio crystals are particularly commonplace in the specialty chemical and pharmaceutical industries as these industries develop complex, high-value-added chemical compounds. Controlling the particle size distribution (PSD) during crystallization of these compounds is critical, both to the efficiency of the downstream manufacturing processes and to the quality of the end-product. However, conventional PSD monitoring technologies, such as laser diffraction and laser backscattering, do not provide the monitoring capability necessary to achieve on-line PSD control for systems in which the particles are highly non-spherical [15]. Laser diffraction overestimates the broadness of the spherical diameter distribution for high-aspect-ratio particles due to orientation effects and the spherical models used to interpret the diffraction data [119, 77]. The chord length distribution measured using laser backscattering can be inverted theoretically to obtain the PSD for high-aspect-ratio particles, but the inversion is highly ill-posed [116] and requires assumptions regarding the particle shape. Successful inversion has been demonstrated experimentally for spheres [50] and octahedra [116], but has been demonstrated for high-aspect-ratio particles using simulations only [63].

Imaging offers the potential to extract both size and shape information and is therefore a promising and attractive approach for obtaining PSD's for non-spherical particles. Several researchers have developed imaging-based methods for sizing elongated crystals [88, 96, 82, 83], but none of these methods are sufficiently automated to be suitable for on-line monitoring and control. Several commercial, imaging-based, on-line particle size and shape analyzers have recently become available, such as Malvern's Sysmex FPIA 3000 and Beckman-Coulter's RapidVUE (see [3] for a survey of other imaging-based instruments). Most of these instruments require sampling, however, which is inconvenient, possibly hazardous, and raises concerns about whether the sample is representative of the bulk slurry [3, 7].One exception is the Mettler Toledo Lasentec Particle Vision and Measurement (PVM) in situ probe. This probe is packaged with automatic image analysis software that, although useful for some systems, does not give suitable results for systems of needle-like particles. The utility of in situ video microscopy has been limited primarily to qualitative monitoring because the nature of in situ images, which contain blurred, out-of-focus, and overlapping particles, has precluded the successful application of image analysis to automatically quantify particle size and shape [15].

The main barrier to successful automation of image analysis for in situ images is "segmen-

tation," that is, separating the objects of interest from the background. Segmentation is simplified if particles are imaged using transmitted light because the particle outlines are easily distinguished [86]. To handle high solids concentrations, however, in situ imaging is constrained to use reflected light, so segmentation is more difficult. [19] presented a technique for automatic segmentation of in-process suspension crystallizer images, but the technique was demonstrated only on images that appear to have been acquired at low solids concentration where there are no overlapping particles and the particles' edges are well-defined. [55] used an in situ fluorescence imaging method that caused the liquid phase of a fluidized bed to fluoresce while leaving the coal particles opaque, thus enabling a gray-level theshold method to detect particle edges. However, for dense particle volume fractions, manual intervention was required to determine which of the segmented particles could be used for sizing.

The goal of this work is to demonstrate robust and efficient segmentation for in situ images of high-aspect-ratio particles using a novel image analysis algorithm. To achieve this goal, we show that the algorithm's PSD measurements are consistent with measurements obtained through manual image analysis by human operators. The accuracy of the measured PSD, therefore, is established only with respect to the PSD measured by human operators. The absolute accuracy of the measured PSD depends on several factors, including the focal depth, the field of view, the solids concentration, the PSD, the particle aspect ratio, and the hydrodynamics of the imaged volume. A complete description of the effects of all of these factors on the absolute accuracy of the measured PSD is beyond the scope of this paper but is a topic of current research in our laboratory.

The chapter is organized as follows. Section 6.2 describes the algorithm, and Section 5.2 presents the experimental studies used to evaluate the algorithm's accuracy and speed. Our findings are summarized in Section 8.6.

5.1 Image Analysis Algorithm Description

This section describes the image analysis algorithm developed for the purpose of analyzing in situ images of high-aspect-ratio crystals. The algorithm is referred to in this paper as SHARC (**S**egmentation for **H**igh-**A**spect-**R**atio **C**rystals) and has been implemented in MATLAB 7.1. We present first an overview of SHARC and then describe in more detail each of SHARC's components.

5.1.1 Overview

The SHARC algorithm is built on the assumption that a needle-shaped crystal can be modeled geometrically as a group of two or more spatially-proximate lines with similar orientation and length. The SHARC algorithm searches for image features satisfying this model in the following manner: First, SHARC detects linear features in the image, referred to as "elementary line segments" or ELSs. Next, SHARC identifies collinear line pairs (lines that appear to belong to a single crystal edge but have been broken up due to background noise, particle overlap, or crystal defects) and creates a representative line, called a "base line," for each pair. Given both the

ELSs and the base lines, SHARC identifies pairs of spatially-proximate, parallel lines of similar length. Finally, SHARC identifies consistent groups of parallel lines and clusters the constituent lines in each of these groups as belonging to a single crystal. The properties (e.g. length, aspect ratio) of these line clusters are used as estimates of the properties of the crystals in the image. Figure S1 shows the result of applying these steps to a small section of an in situ image of needle-like pharmaceutical crystals.

5.1.2 Linear Feature Detection.

Line segments are commonly used as inputs to higher-level processes in machine vision, and many different methods have been developed for extracting line segments from images (see [53] for a review of these methods). SHARC uses the Burns line finder [17]. For our application, the Burns line finder is advantageous over the popular Hough transform-based methods for several reasons. First, the Burns line finder is scale-independent, that is, it finds short lines just as easily as it finds long lines. The Burns line finder also has lower computation and memory requirements than the Hough transform and finds line endpoints more easily. The Burns line finder is unique in that it detects lines on the basis of image intensity gradient *direction*, whereas most line-finders are based on image intensity gradient *magnitude*. The Burns line finder is therefore able to detect subtle linear features that would be missed by other line finders. This feature also means that its performance is relatively insensitive to variations in contrast and brightness. This property is important for crystallization imaging because, as crystallization occurs, the increasing solids concentration causes more reflected light to reach the camera CCD, resulting in image intensity variations. Such variations do not affect the performance of the Burns line finder.

We have modified slightly the Burns algorithm to enhance its performance for our particular application, incorporating some of the speed-up suggestions given by [54]. Our implementation consists of the following steps:

1. Calculate the direction and magnitude of the image intensity gradient at each pixel using a Sobel gradient operator of size $n_\nabla \times n_\nabla$.

2. For each pixel with gradient magnitude above a small threshold $\epsilon_{|\nabla|}$, assign a gradient direction label by coarsely quantizing the pixel's gradient direction into one of n_b sets of ranges, or "buckets," as depicted in Figure S2.

3. Apply a connected components algorithm (CCA) to group identically-labeled, adjacent (including diagonally adjacent) pixels into "line support regions," as depicted in Figure S3.

4. Filter line support regions that have a pixel area less than some pre-defined threshold ϵ_A.

5. Fit a line to each remaining line support region, as depicted in Figure S3.

To eliminate quantization effects, steps 2–4 are performed twice before proceeding to step 5, each time using a different quantization in which the gradient direction partitioning is shifted by half the bucket size, as shown in Figure S2. This procedure results in each pixel being associated

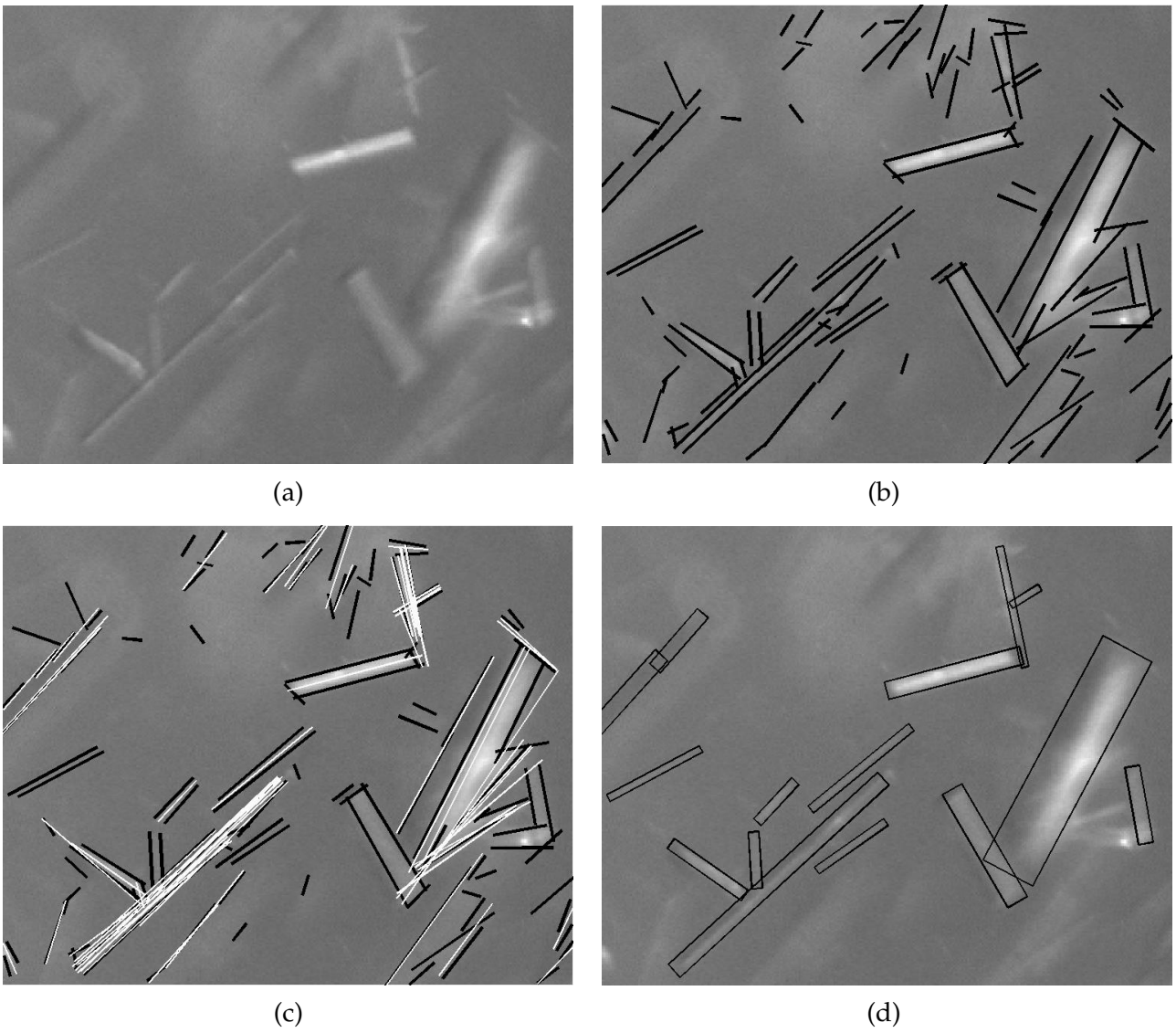


Figure S1: Example of SHARC algorithm applied to an in situ image of suspended pharmaceutical crystals. (a) A region of interest in the original image. (b) Linear features (ELSs) extracted from the original image. (c) ELSs (black lines) and lines representing each collinear line pair (white lines). (d) Representative rectangles for clusters of spatially-proximate parallel lines with roughly equal length. The lengths, widths, and aspect ratios of the rectangles are used as the crystal size and shape measurements.

with two different line support regions, and this conflict is resolved through a voting process designed to select the interpretation that results in the longest possible line support regions [17]. To reduce computation time, SHARC carries out this voting process on the basis of the pixel areas of the conflicting line support regions, which in almost every case gives the same results as voting

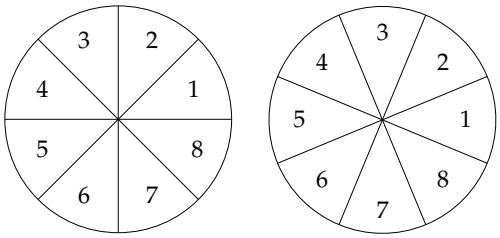


Figure S2: Depiction of different eight-bucket gradient direction quantizations used to label pixels. For the quantization on the left, pixels having gradient direction in the range of 0 to 45 degrees are labeled as "1", pixels with gradient direction in the range of 45 to 90 degrees are labeled as "2", and so forth. Quantization effects are mitigated by applying a second quantization, such as that shown on the right, and subsequently resolving any conflicts between the results given by each quantization.

based on length.

The line-fitting method used in step 5 is standard blob analysis, available in most image analysis packages. Blob analysis fits a line to each region of pixels by determining an ellipse having the same geometric moments as the region, as depicted in Figure S3. The ellipse's major axis length, minor axis length, orientation, and centroid are used respectively as the length, width, orientation, and center of the corresponding line.

5.1.3 Identification of Collinear Line Pairs

During linear feature extraction, single edges are commonly broken up into multiple collinear lines. This problem is common for systems of needle-like crystals because the particles are often touching or overlapping. Therefore, SHARC groups these collinear lines prior to searching for groups of parallel lines having similar length and orientation. The problem of collinear line grouping has been studied extensively. [53] compare and evaluate a number of the available methods. SHARC uses a straightforward, computationally inexpensive method developed by [31]. Etemadi's method involves projecting ELSs of similar orientation onto a common line to determine if the lines satisfy simple spatial-proximity thresholds given by

$$|\theta_1 - \theta_2| < \epsilon_{\theta_C}, \quad d_{\text{PD}} < \epsilon_{\text{PD}}(w_1 + w_2), \quad d_{\text{EP}} < \epsilon_{\text{EP}}(L_1^P + L_2^P) \tag{5.1}$$

in which d_{PD} is the perpendicular distance between the two lines and d_{EP} is the distance between their nearest projected endpoints. θ_i and w_i are, respectively, the orientation and width of line i calculated using blob analysis, and L_i^P is the projected length of line i, calculated as described below. ϵ_{θ_C}, ϵ_{PD}, and ϵ_{EP} are user-specified thresholds.

The perpendicular and endpoint distances and projected lengths are calculated, as depicted in Figure S4, by projecting the two lines onto a "virtual line" whose position (x_V, y_V) and orientation (θ_V) are length-weighted averages of the positions and orientations of the constituent ELSs,

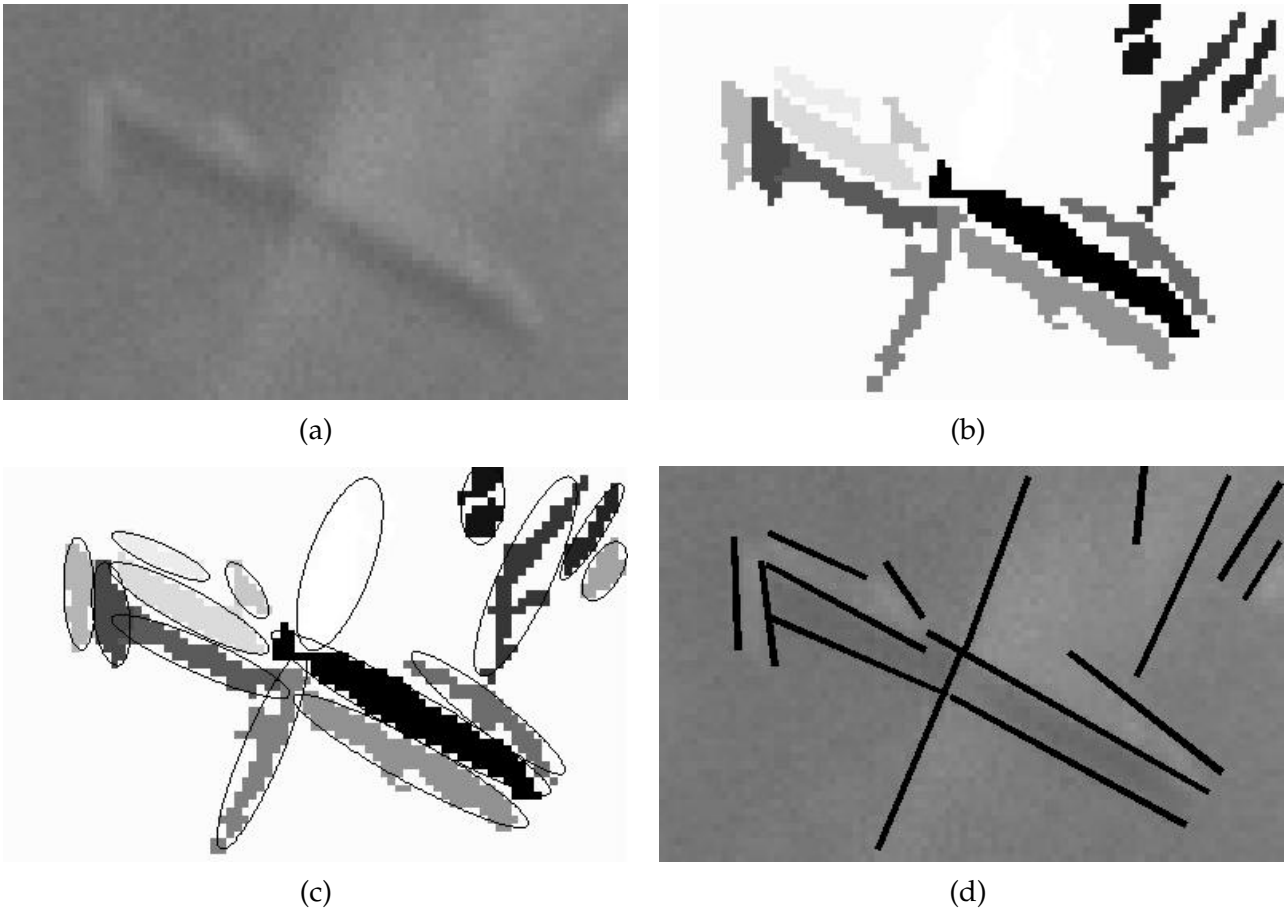


Figure S3: Example of finding linear features using Burns line finder and blob analysis. (a) Grayscale image. (b) Regions of pixels having similar gradient orientation, determined using the Burns line finder. (c) Best-fit ellipses for each region of pixels, determined using blob analysis. (d) Major axes of the best-fit ellipses imposed on the original grayscale image.

given by

$$x_V = \frac{L_1 x_1 + L_2 x_2}{L_1 + L_2}, \quad y_V = \frac{L_1 y_1 + L_2 y_2}{L_1 + L_2}, \quad \theta_V = \frac{L_1 \theta_1 + L_2 \theta_2}{L_1 + L_2} \tag{5.2}$$

in which L_i, θ_i, x_i, and y_i are the length, orientation, horizontal centroid, and vertical centroid of ELS i, calculated using blob analysis. Given the position and orientation of the virtual line, the perpendicular distance between the two ELSs can be calculated as the sum of the perpendicular distances of the ELSs centroids from the virtual line. The length of the virtual line, L_V, is defined as the length of the shortest possible line containing all four projected endpoints.

For each line pair satisfying the collinearity criteria, the corresponding virtual line becomes a base line and is subsequently used in the identification of parallel pairs, as described in the following section.

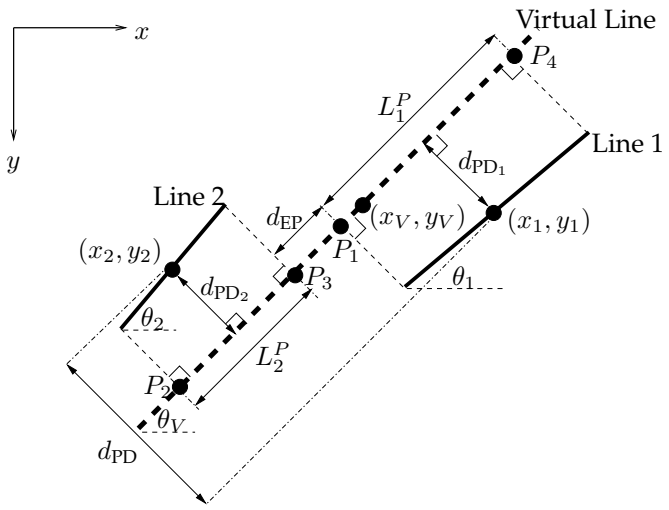


Figure S4: Depiction of variables used in line pair classification scheme.

5.1.4 Identification of Parallel Line Pairs

Following collinear line pair identification, SHARC identifies pairs of parallel lines, or lines that have similar orientation, are spatially proximate, and exhibit a high degree of overlap when projected onto a common line. These line pairs satisfy the following criteria:

$$|\theta_1 - \theta_2| < \epsilon_{\theta_P}, \quad d_{\text{PD}} < \frac{1}{\epsilon_{\text{AR}}} L_{max}, \quad Q_P > \epsilon_Q \tag{5.3}$$

in which ϵ_{θ_P}, ϵ_{AR}, and ϵ_Q are user-specified thresholds for orientation difference, aspect ratio, and pair "quality." d_{PD} is the perpendicular distance between the two lines, and L_V is the length of the virtual line, as defined in the previous section. L_{max} is the length of the longest line in the pair, and Q_P quantifies the "quality" of the pair. The quality metric used by SHARC, and suggested by [31], is based on the degree of overlap of the two parallel lines, calculated using

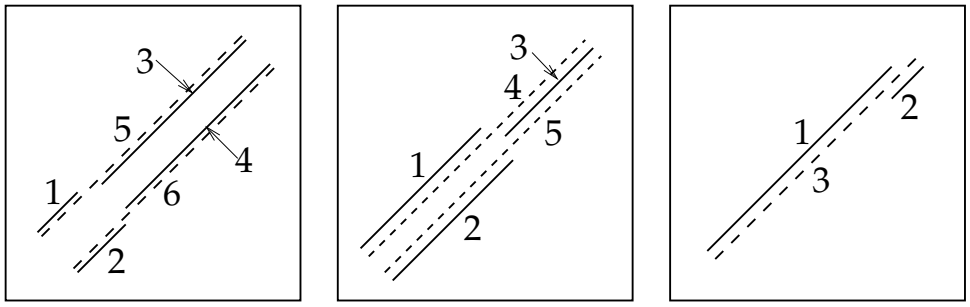
$$Q_P = \frac{L_1^P + L_2^P}{2L_V} \tag{5.4}$$

in which the projected lengths L_i^P are computed as described in Section 5.1.3. This metric is simple to compute and scale-independent, depending only on the relative lengths of the lines. Overlapping parallel pairs give a Q_P between 0.5 and 1, the latter value representing a perfectly overlapping pair.

If the parallel pair includes a base line comprising two collinear ELSs, it is possible that the two lines in the parallel pair share an ELS, in which case the pair is invalid and is discarded. Figure S5 depicts examples of valid and invalid parallel pairs.

Each parallel pair is ranked according to its significance, calculated as

$$S = \frac{L_{min}^2}{L_{max}} \tag{5.5}$$



(a) (b) (c)

Figure S5: Examples of valid and invalid parallel line pairs. The solid lines represent ELSs and the dotted lines represent base lines (lines arising from instances of collinearity). In (a), the base lines 5 and 6 form a valid parallel pair, and the ELSs 3 and 4 also form a valid parallel pair. In (b), the parallel lines 4 and 5 are an invalid parallel pair because they both depend upon ELS 3. Similarly, in (c), base line 3 and ELS 1 form an invalid pair because both depend on ELS 1.

in which S is the significance level and L_{min} and L_{max} are, respectively, the lengths of the shorter and longer lines in the pair. This significance measure is used to account for the fact that longer lines are less likely to have arisen by accident or due to noise and should thus be considered more significant. The significance ranking is used to order the subsequent line clustering procedure but affects the results only when there is a conflict between two high quality pairs, or when two high quality pairs are mutually exclusive. These conflicts arise since SHARC identifies parallel pairs using both the ELSs and the base lines (for example, ELSs 3 and 4 in Figure S5(a) form a valid parallel pair but are also involved indirectly in the parallel pair of base lines 5 and 6). If a given ELS is involved in two conflicting parallel pairs, the significance ranking is used in the line clustering process to favor the interpretation that leads to the longer crystal.

5.1.5 Clustering

The objective of clustering is to group those lines that appear to belong to a single crystal. The clusters are formed by (1) identifying the most significant parallel pair on the basis of equation (5.5), (2) recursively identifying all other lines that are parallel-paired with at least one of the lines in the current cluster, and (3) removing from the list of parallel pairs any pairs that include an ELS or base line associated with the newly-formed group. This process is iterated until the list of parallel pairs is empty, after which all lines that are not included in any of the formed groups are discarded.

The properties of each line cluster are calculated using the method described in Section 5.1.3, generalized to an arbitrary number of lines. That is, the cluster orientation is the length-weighted average of all lines in the cluster, the cluster length is the length of the shortest possible line containing all projected endpoints for all lines in the cluster, and the cluster width is the largest possible perpendicular distance between the centroids of all lines in the cluster. Clusters having an aspect ratio below the user-defined threshold ϵ_{AR} are discarded.

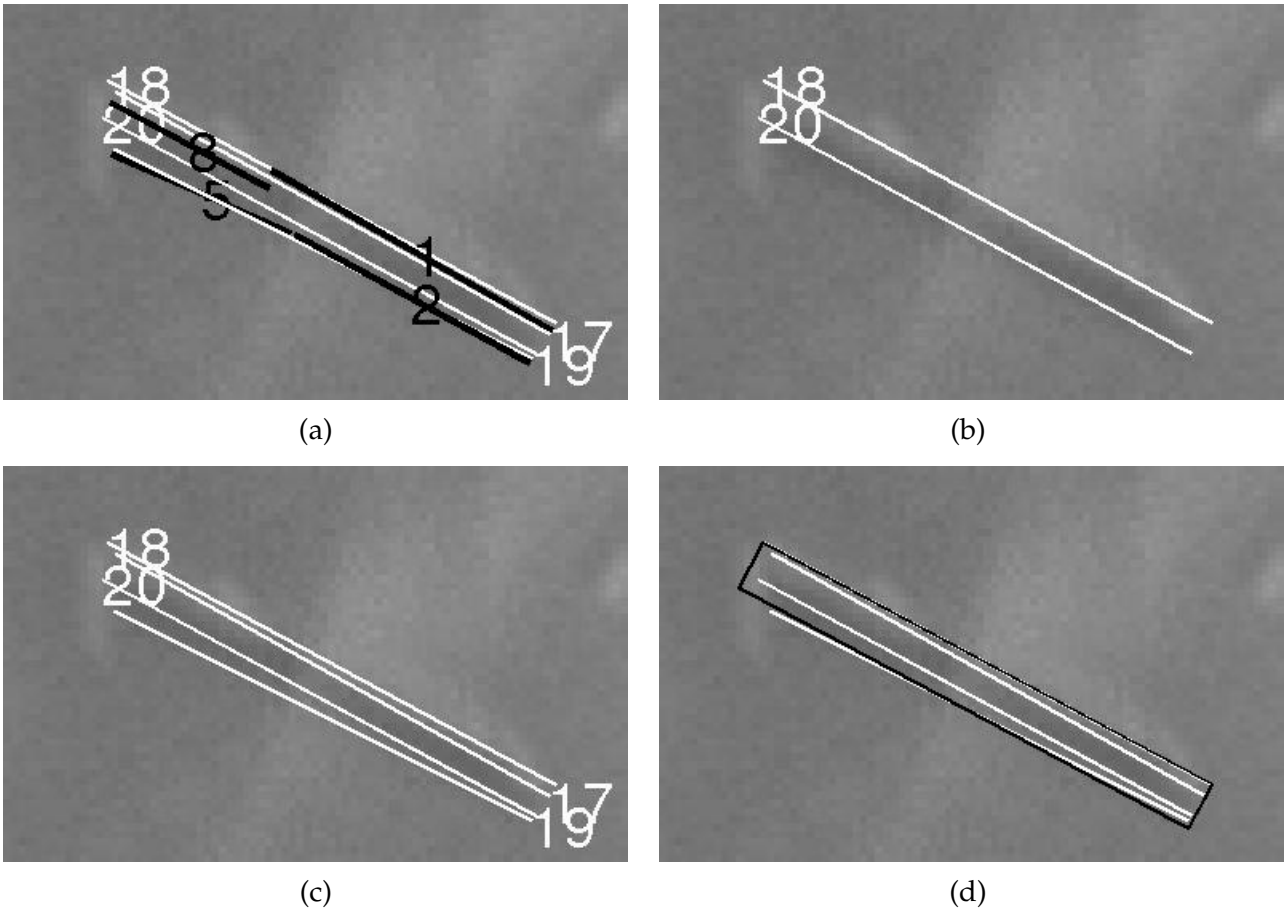


Figure S6: Example of clustering procedure for valid parallel pairs. (a) ELSs (dark) and base lines (light) involved in at least one valid parallel pair. (b) The pair with the highest significance. (c) Lines that are parallel-paired (either directly or indirectly) with either of the lines in the highest significance pair. (d) The bounding box calculated for the line cluster.

Figure S6 illustrates the clustering procedure using the set of lines extracted in Figure S3. Figure S6(a) shows all lines involved in at least one valid parallel pair, including both ELSs and base lines. The valid parallel pairs for this example are (10,14), (12,15), (18,19), (18,20), and (19,21). Lines 18, 19, 20, and 21 are base lines comprising the collinear pairs (9,15),(10,12),(12,14), and (14,15), respectively. Figure S6(b) shows that pair (18,20) has the highest significance and is therefore analyzed first in the clustering order. Figure S6(c) shows the result of recursively identifying all lines that are paired with lines in the cluster. That is, line 19 is identified in the first recursion due to its pairing with line 18, and line 21 is identified in the second recursion due to its pairing with line 19. The rectangle in Figure S6(d) indicates the length, width, and orientation of the line cluster. The parallel pairs (10,14) and (12,15), each of which has at least one of its members involved in the newly-formed grouping, are removed from the list of parallel pairs.

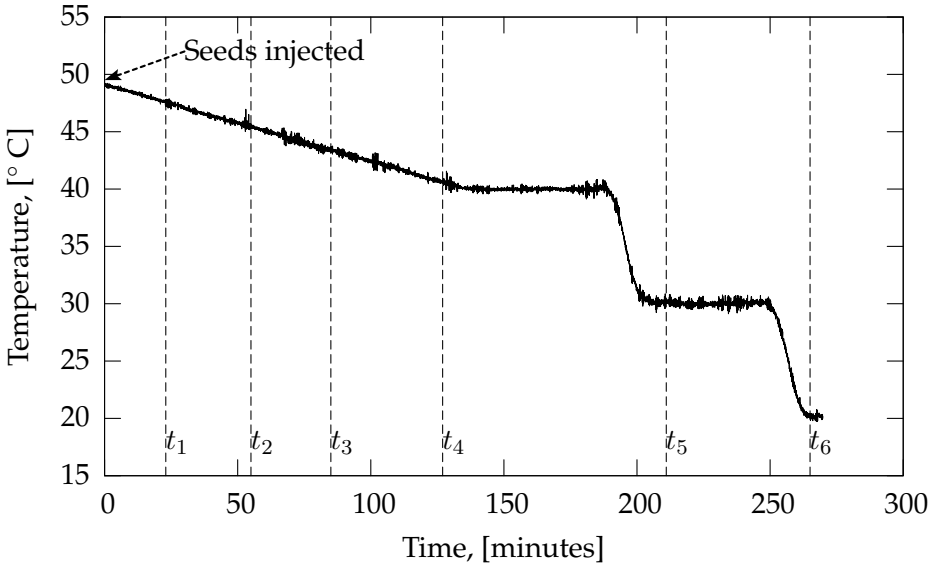


Figure S7: Temperature trajectory for crystallization experiment. The vertical lines indicate the times at which sets of video images were acquired.

Line finder Parameters		Collinearity Thresholds		Parallelism Thresholds	
n_∇	5	ϵ_{θ_C}	20 degrees	ϵ_{θ_P}	5 degrees
$\epsilon_{\lvert\nabla\rvert}$	1	ϵ_{EP}	0.5	ϵ_Q	0.85
n_b	6 buckets	ϵ_{PD}	0.5	ϵ_{AR}	4.5
ϵ_A	20 pixels				

Table 5.1: Parameter values used to analyze images from crystallization experiment.

5.2 Experimental Results

To evaluate SHARC's performance, a seeded, pharmaceutical crystallization was carried out during which several sets of video images were acquired. The images in each set were acquired over a few seconds only, such that we assume the properties of the crystal population are constant for each image set. The temperature profile (following seed injection) for this crystallization is shown in Figure S7. The time at which each set of video images was acquired is indicated in Figure S7 by a vertical line and labeled as t_j, the subscript indicating the image set number. The mixing speed for this experiment was 550 RPM (1.1 m/s tip speed), sufficiently fast that the crystals could not be tracked from one video frame to the next. The same parameter values were used to analyze all images (see Table 8.2).

The following sections assess SHARC's suitability for on-line monitoring and control with respect to both accuracy and speed.

5.2.1 Algorithm Accuracy

Visual Evaluation

Figures S8–S11 show SHARC's performance on selected images taken from video sets 3, 4, 5, and 6. These figures demonstrate SHARC's effectiveness for images having poor contrast and uneven background. These figures also demonstrate SHARC's ability to detect crystals with poorly-defined edges, varying intensity levels, and a certain level of particle overlap.

Comparisons with Manual Sizing

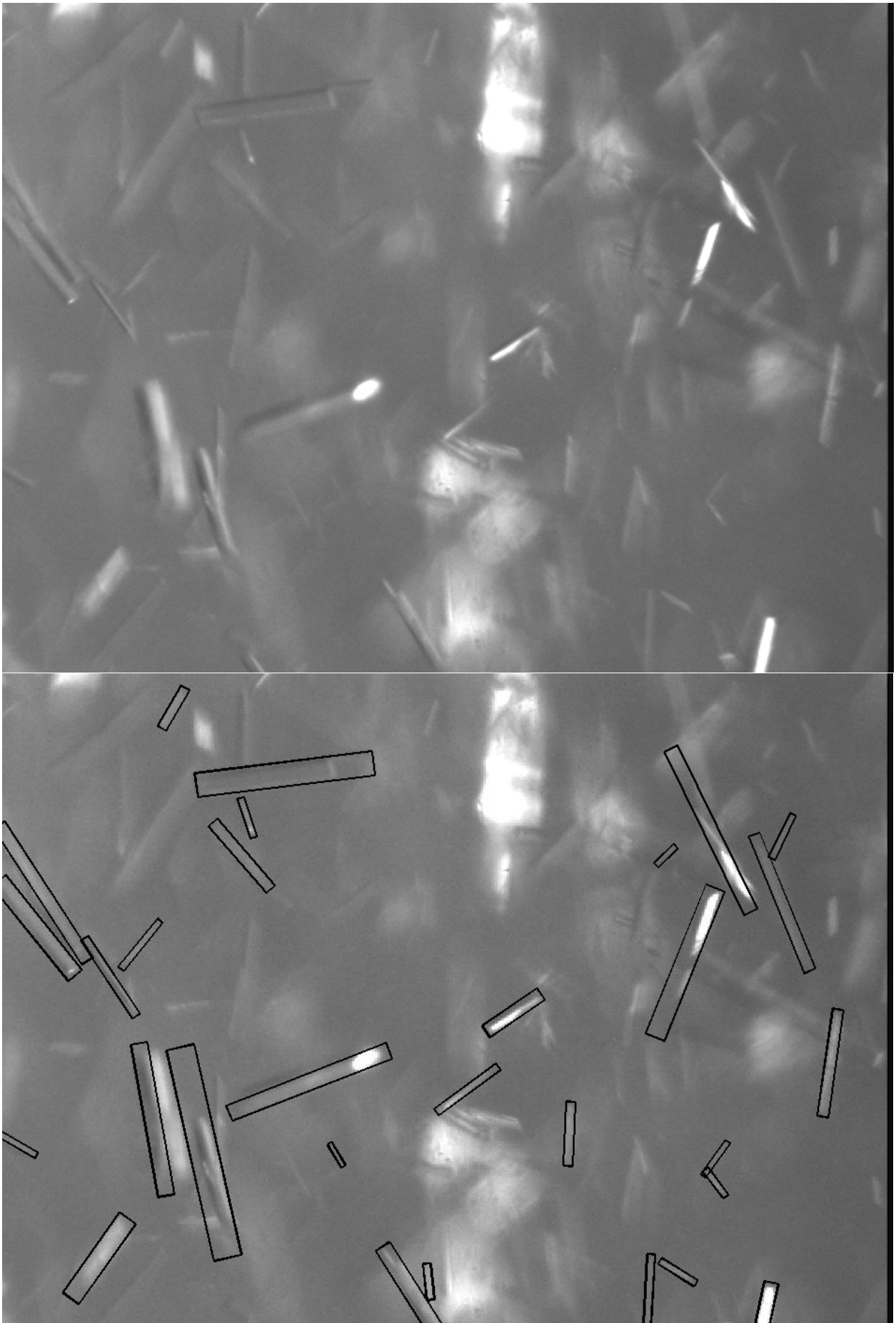
We evaluate SHARC's accuracy by comparing its PSD measurements with measurements obtained through manual image analysis by human operators. Although the human vision system is clearly more reliable than current computer vision systems, manual image analysis introduces an undesirable level of subjectiveness into the measurement. The subjectiveness of manual sizing is magnified for in situ images because the crystals often appear blurry, overlapping, and out-of-focus. Thus, it can be difficult to decide whether or not a given crystal is sufficiently in focus and well-defined to be sized accurately. To assess the subjectiveness involved in manually sizing in situ images and determine what constitutes "good" agreement between SHARC's measurement and a manual measurement, we asked nine different people to manually size crystals for the same ten images from image set 3. We confirmed that ten images was sufficient to achieve convergence of the overall mean size measurement for each operator. Table **??** shows the overall mean size calculated for all ten images for each operator, the total number of crystals found by each operator, and the average percent difference in overall mean size between each operator and the other operators. This latter value is calculated for operator i using the equation

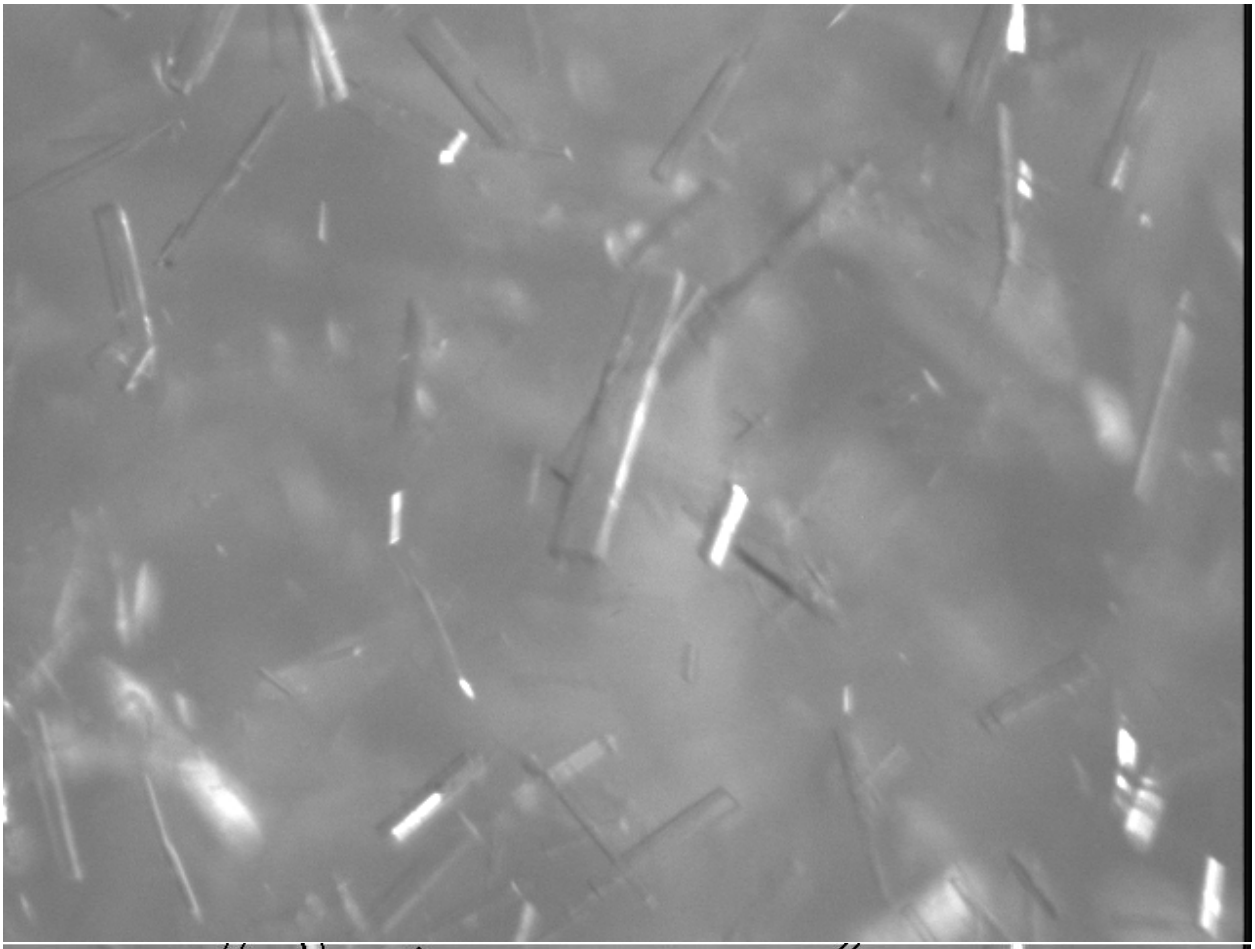
$$\text{diff}_i = \frac{100}{N_{\text{op}} - 1} \sum_{j=1}^{N_{\text{op}}} \frac{|x_i - x_j|}{\frac{x_i + x_j}{2}} \tag{5.6}$$

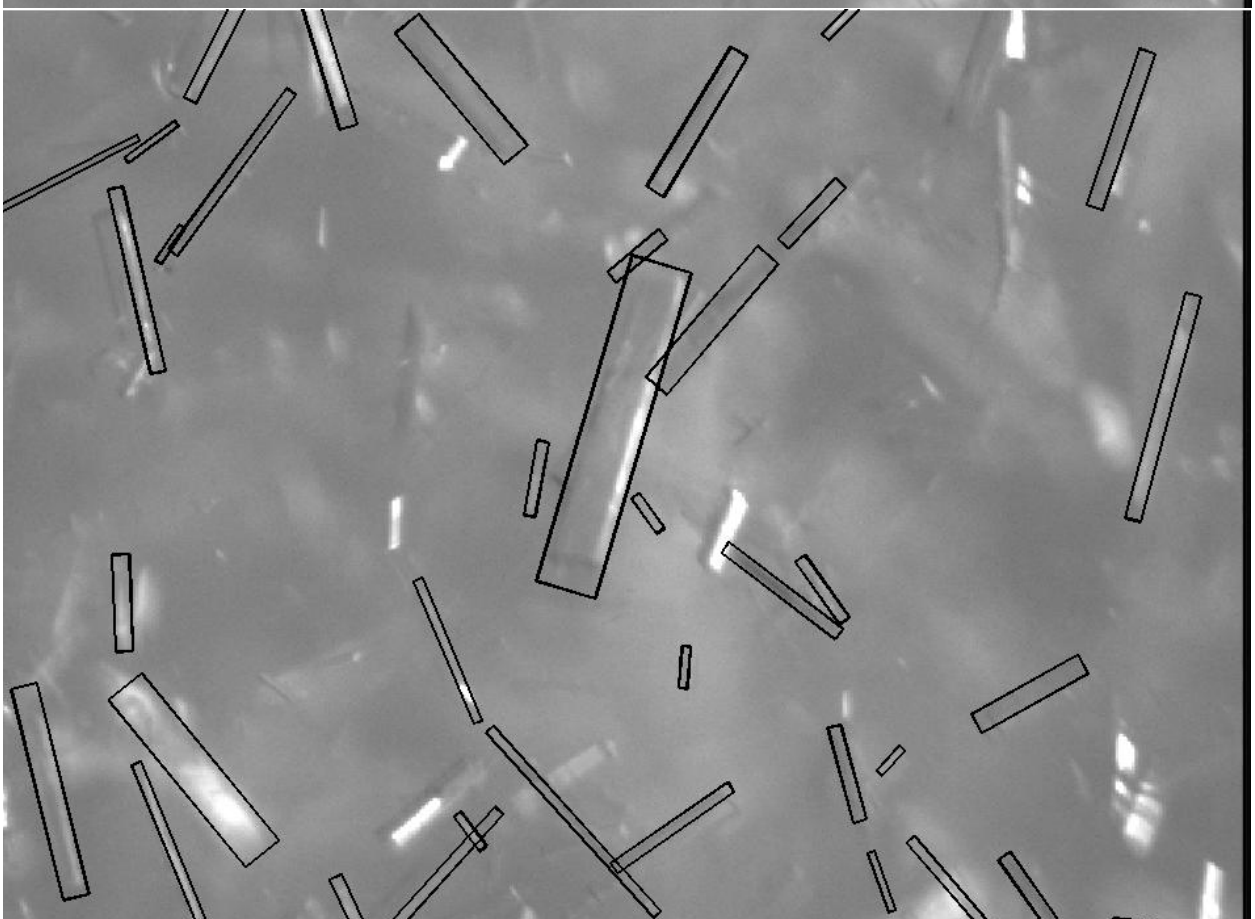
Table **??** shows that the mean size varied by as much as 37% between operators, and the number of crystals found varied by over 100% between operators, illustrating the large degree of subjectiveness associated with manual sizing of in situ images. However, the relatively small standard deviation in overall mean size indicates that manual sizing constitutes a reasonably reliable standard provided the measurement is performed by a sufficient number of operators.

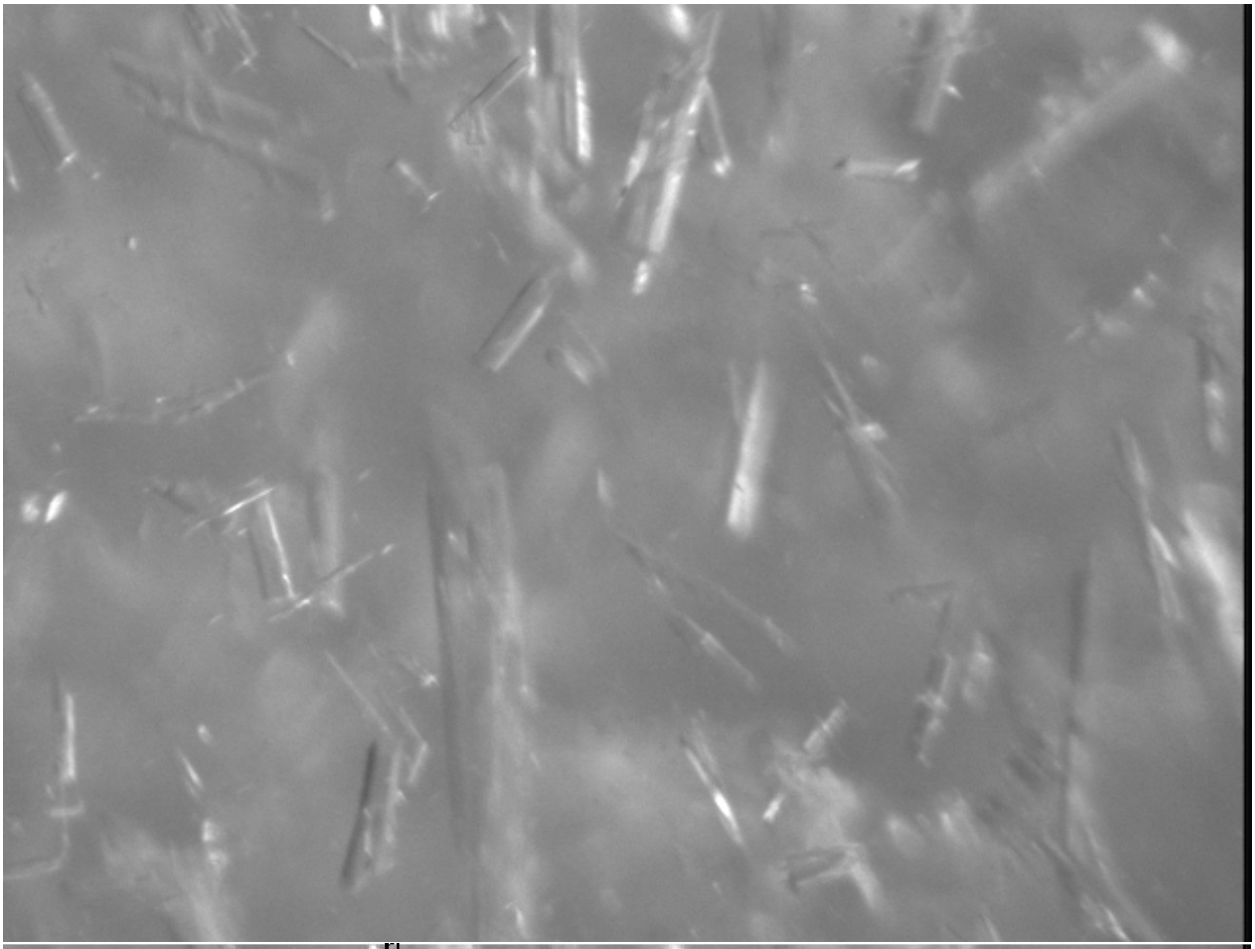
Based on the results shown in Table **??**, we define "good" agreement between SHARC's results and a set of manual results to mean that their individual means are within approximately 12% of their combined mean value. Table 5.2.1 shows that the SHARC algorithm determines a mean crystal size within 2% of that found by the nine manual operators, and Figure S12 shows a good match between the cumulative distribution function found by SHARC and the cumulative distribution functions of the nine manual operators.

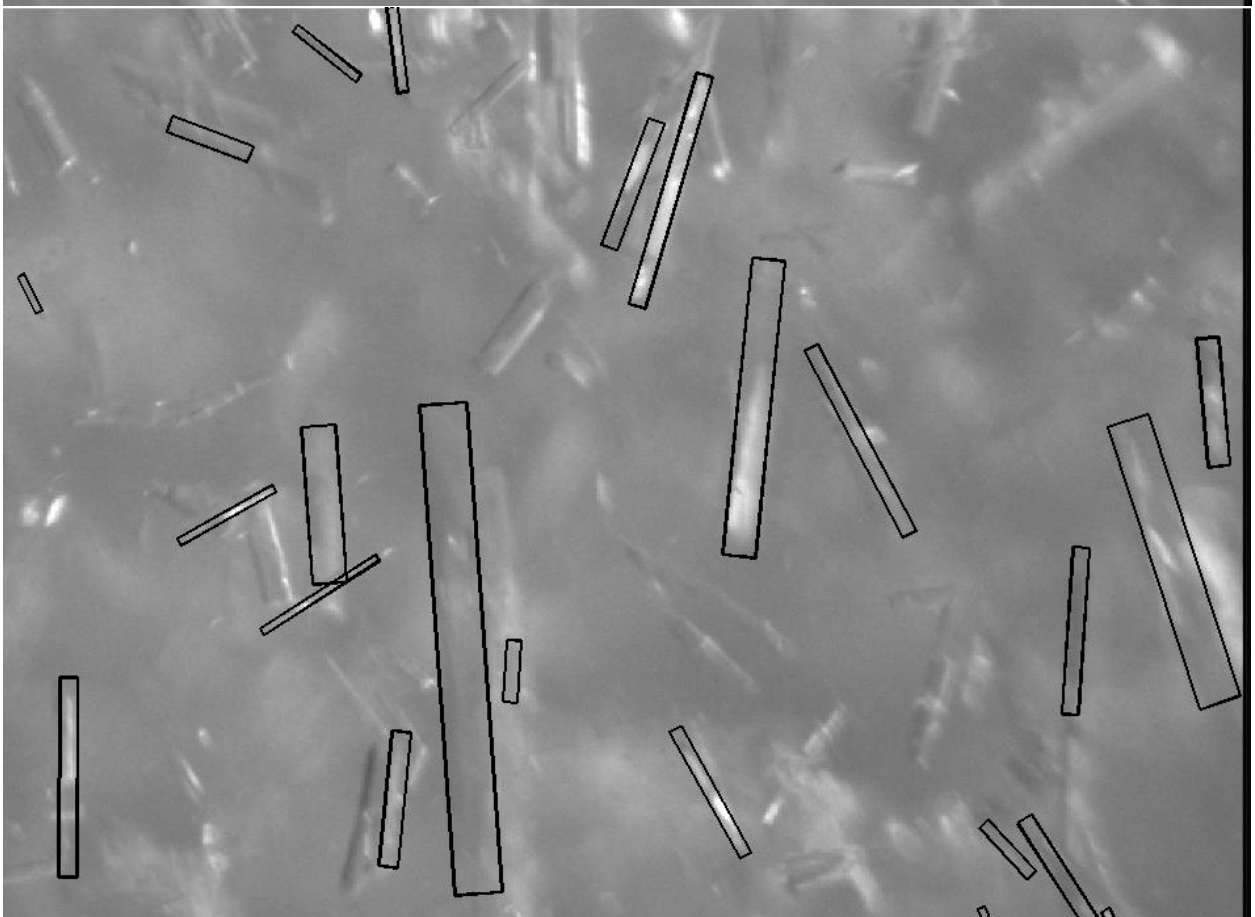
To evaluate SHARC's ability to maintain accuracy for the duration of the experiment, twenty-five images from sets 4, 5, and 6 were analyzed both manually (by a single operator) and using SHARC. Table 5.2.1 shows that SHARC maintains good agreement with the results

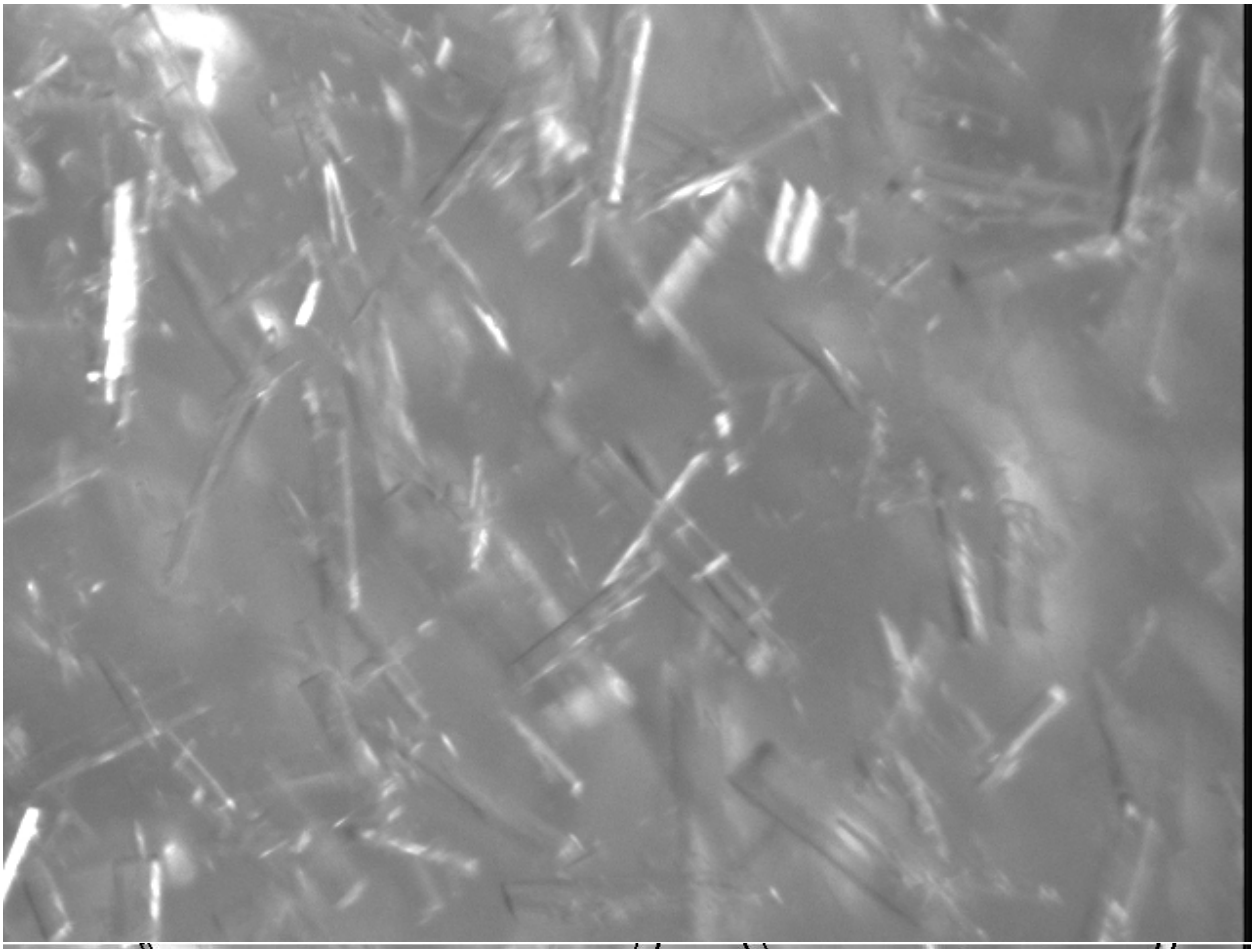


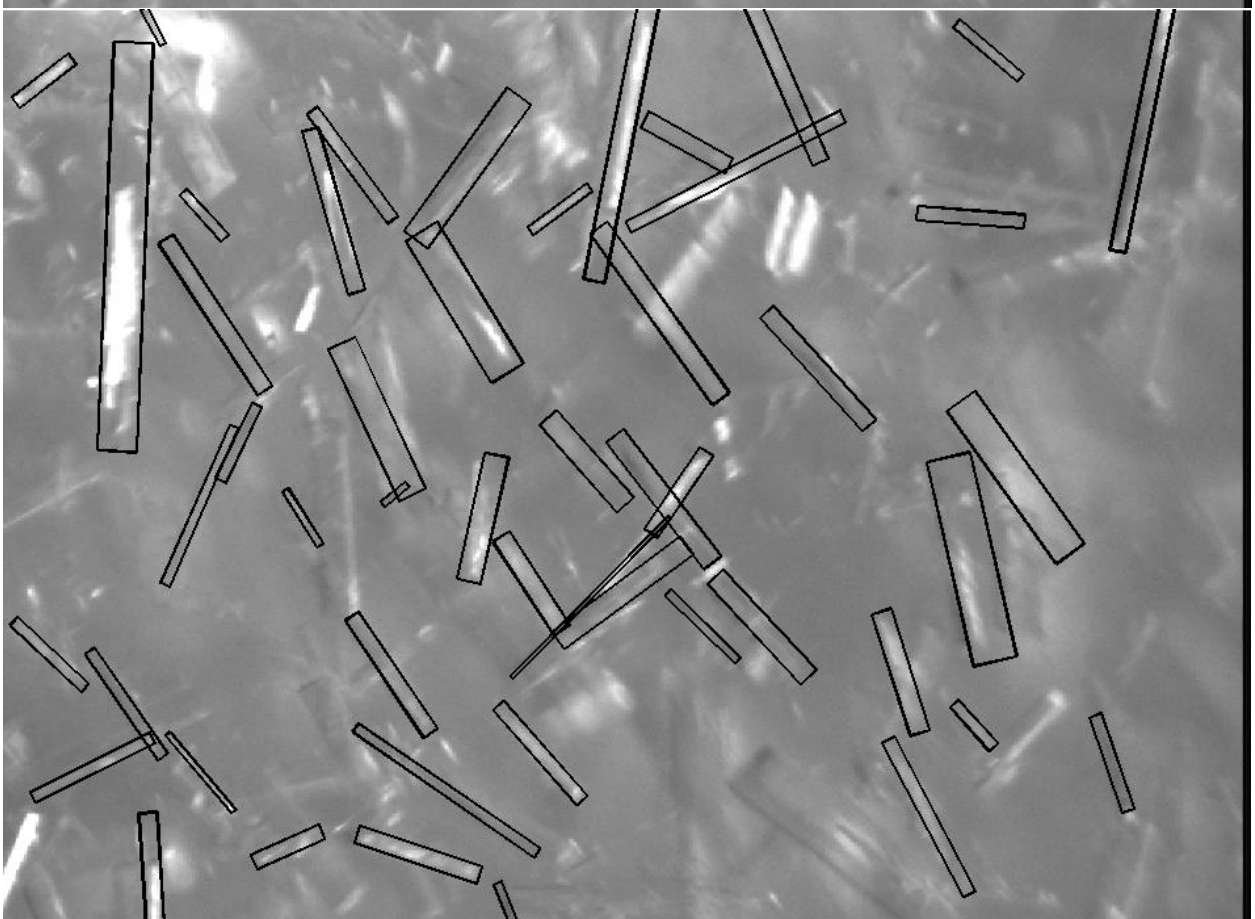












	Mean Size, [μm]		
Image set	Manual	Automatic	% Difference
3	283	284	0.3
4	290	286	1.2
5	306	278	9.3
6	355	283	22.5

Table 5.2: Comparison of mean sizes calculated from manual sizing of crystals and from automatic sizing.

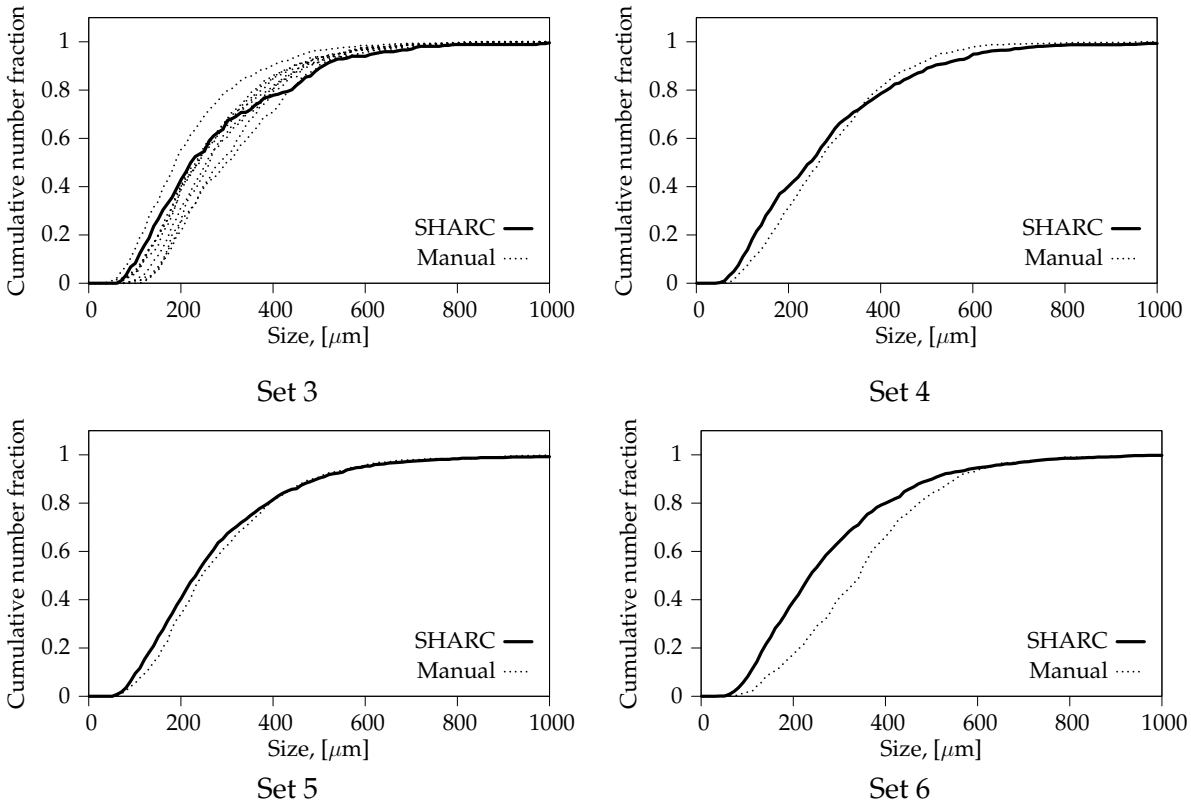


Figure S12: Comparison of cumulative number fractions obtained from manual and automatic sizing of crystals for video image sets 3, 4, 5, and 6. Set 3 was manually sized by nine different operators.

obtained manually for sets 4 and 5, but its performance declines somewhat for set 6. These same conclusions can be drawn from Figure S12, which compares the cumulative distribution functions obtained using both methods.

Figure S12 and Table 5.2.1 indicate that, as the solids concentration and degree of crystal attrition increase, SHARC either fails to identify a significant percentage of the larger crystals or erroneously identifies smaller crystals. Figure S13 shows the results of manually sizing an

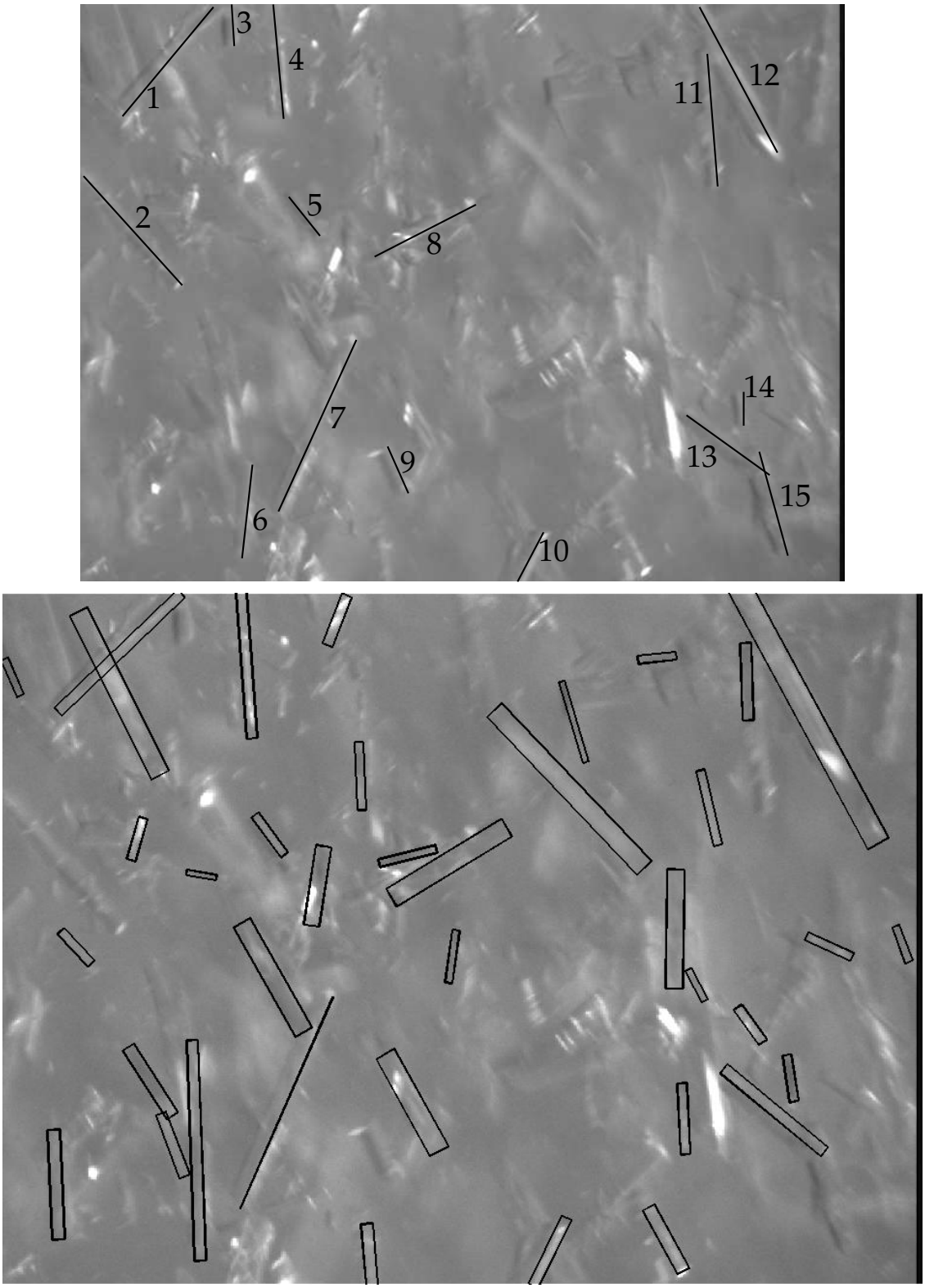


Figure S13: Comparison of crystals sized manually (top) and using SHARC (bottom).

Set	average total cputime per image [s]	average cputime per image [s] linear feature detection	collinearity identification	parallelism identification	clustering
2	1.9	1.8 (97)	0.0 (1)	0.0 (0)	0.0 (0)
3	3.1	2.0 (65)	0.4 (14)	0.6 (19)	0.0 (0)
4	4.3	2.1 (47)	0.8 (17)	1.5 (33)	0.0 (0)
5	7.1	2.5 (35)	1.4 (19)	3.2 (44)	0.0 (0)
6	10.8	2.9 (26)	2.2 (20)	5.6 (52)	0.1 (0)

Table 5.3: Computational requirements for analyzing different image sets (averaged over 10 images). The numbers in parentheses give the percentages of total cputime. Images are analyzed using a 2.2 GHz AMD Athlon 64 processor.

image from set 6 compared with the results given by SHARC. SHARC identifies several smaller crystals not identified manually as well as only portions of some of the bigger crystals. These misidentifications explain why SHARC's mean size is less than the mean size obtained manually.

Some of the differences between SHARC's results and the manual results can be attributed to the subjectiveness associated with manual sizing. However, crystals 2, 11, and 15 are clearly misidentifications. Figure S14 shows a zoomed-in view of these crystals and demonstrates the results of each step in SHARC. For these cases, particle overlap and attrition interferes with the detection of the crystal edges to such an extent that SHARC is unable to detect important instances of collinearity. These particular misidentifications can be corrected by relaxing SHARC's collinearity criteria, but this would likely lead to further false positives. Investigating the performance of the many available collinear identification methods noted by [53] may be the best way to improve algorithm performance for high solids concentrations.

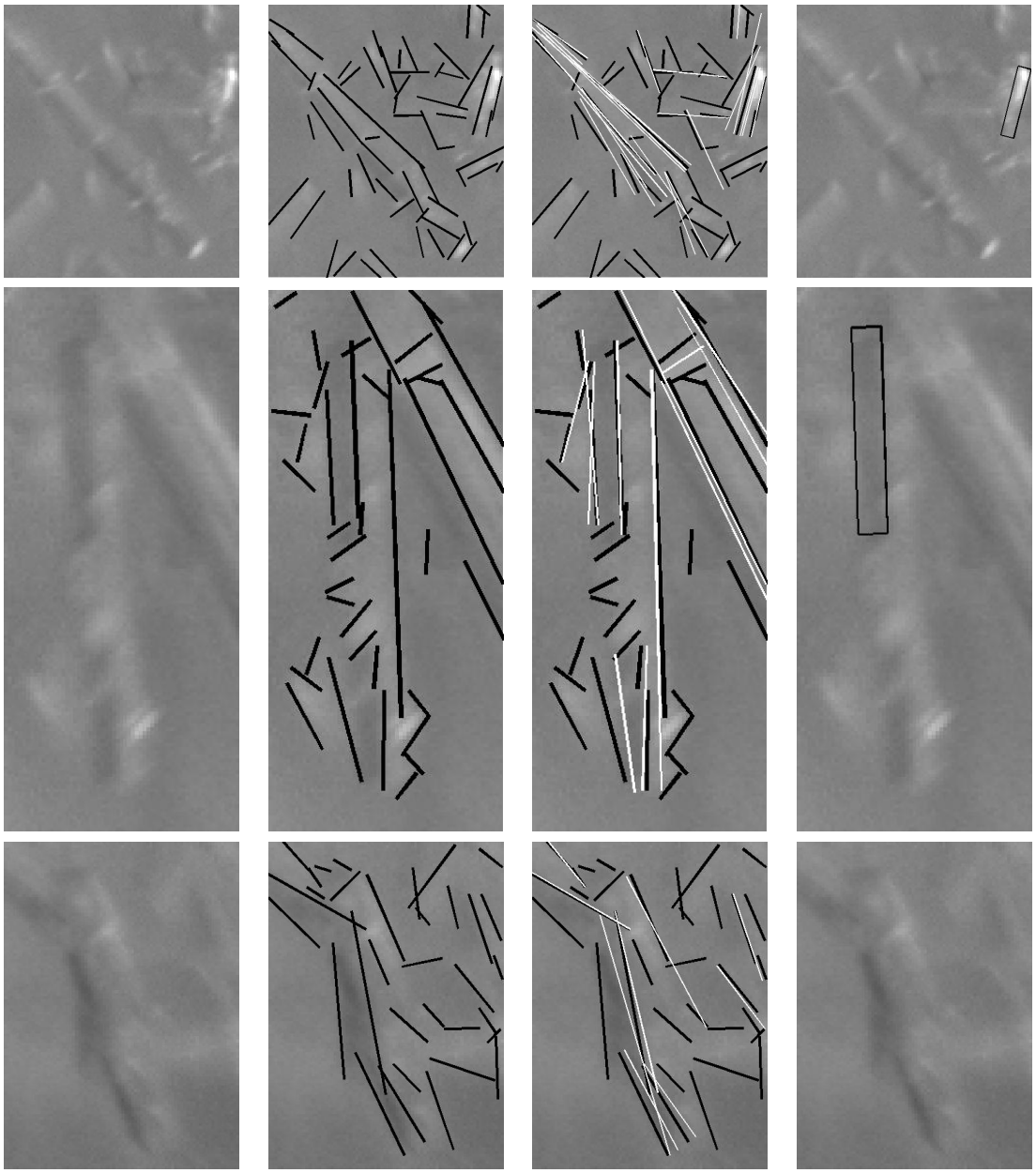
5.2.2 Algorithm speed

This section assesses whether SHARC is sufficiently fast to be useful for on-line monitoring and control of crystallization processes. Table 5.2.2 shows the amount of time required to process an image for each of the image sets and shows how this time is partitioned amongst the different steps of SHARC. Table 5.2.2 indicates that a significant part of the processing time for most sets is spent on the collinear and parallel grouping operations. These operations can be made more efficient by limiting the computations to line pairs that are spatially proximate using a data structure in which the lines are sorted by their endpoint locations, as suggested by [65].

To determine how much time is required to obtain a sufficiently accurate estimate of the mean particle length, we find the number of samples N such that the size of the 95% confidence interval for the population mean particle length is less than 10% of the sample mean. That is, we find the smallest N for which

$$t(\alpha, N-1)\frac{s_N}{\sqrt{N}} \leq 0.1 l_N \tag{5.7}$$

in which t is the t-distribution, α is the confidence level, l_N is the sample mean, and s_N is the



(a) (b) (c) (d)

Figure S14: Zoomed-in view of crystals that SHARC failed to identify correctly. From top to bottom, the crystal labels are two, eleven, and fifteen (according to labels in Figure S13). Column (a): Original image. Column (b): ELS data. Column (c): ELSs and base lines. Column (d): Result of clustering.

Set	average # of crystals per image	# of crystals to converge to mean	cputime to converge to mean [min.]	cputime to converge to variance [min.]
2	1	93	1.9	18.5
3	25	163	0.3	1.8
4	28	129	0.3	2.2
5	30	106	0.4	3.4
6	38	143	0.7	4.2

Table 5.4: Computational requirements to achieve convergence of mean and variance.

sample standard deviation, defined as

$$l_N = \frac{1}{N}\sum_{i=1}^{N} l_i \tag{5.8}$$

$$s_N = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(l_i - l_N)^2} \tag{5.9}$$

in which l_i is the length of particle i. Similarly, to determine how much time is required to obtain a sufficiently accurate estimate of the variance in particle length, we find N such that the size of the 95% confidence interval for the population variance is less than 10% of the sample variance s_N^2, satisfying

$$\frac{N-1}{\chi^2(\alpha, N-1)} - 1 \leq 0.1 \tag{5.10}$$

in which χ^2 is the chi-squared distribution (see [101, p.75]). Equation 5.10 is satisfied for $N = 889$ samples.

Given SHARC's speed, the number of crystals per image, and the number of samples necessary to obtain sufficient measurement accuracy, we can calculate the rate at which SHARC provides accurate mean and variance measurements. Table 5.4 indicates that SHARC requires approximately two minutes to measure the PSD mean and fifteen minutes to measure the PSD variance. Given the time scales of most crystallization processes, SHARC is sufficiently fast to provide measurements for a feedback control system based on measurements of the PSD mean and variance. If necessary, the measurement rate can be increased by implementing SHARC in a compiled language.

5.3 Conclusion

The SHARC algorithm can robustly and efficiently extract crystal size information from in situ images of suspended, high-aspect-ratio crystals for moderate solids concentrations, giving results consistent with measurements obtained through manual image analysis by human operators. SHARC's performance declines for high solids concentrations and high levels of particle

attrition because the degree of particle overlap and the noise arising from attrited particulate matter hinder the identification of the suspended crystals' edges. Implementing improved methods for identifying instances of collinearity may enable suitable performance for these conditions.

The speed with which SHARC analyzes the images is suitable for real-time monitoring and control of PSD mean and variance.

We gratefully acknowledge GlaxoSmithKline for providing the imaging system used in this study. We also gratefully acknowledge the financial support of the industrial members of the Texas-Wisconsin Modeling and Control Consortium. Finally, we wish to thank Joshua, Jeremiah, and Mary Diaz, Aaron Lowe, Murali Rajamani, Ethan Mastny, Michael Benton, and Thomas Knotts, who willingly performed tedious manual sizing of the imaged crystals.

Chapter 6

Three-dimensional Object Recognition for Complex Crystal Shapes

Solution crystallization processes are widely used in the pharmaceutical, food, and chemical industries to create solids with desirable properties and separate or purify chemical species. Controlling the particle size distribution (PSD) and the distribution of particle shapes during crystallization is critical, both to the efficiency of the downstream manufacturing processes and to the quality of the end-product. However, conventional PSD monitoring technologies, such as laser diffraction and laser backscattering, are based on assumptions of particle sphericity and thus do not provide the monitoring capability necessary to achieve on-line PSD control for systems in which the particles are highly non-spherical [119, 15]. Although methods have been developed for inverting the chord length distribution (CLD) measured by laser backscattering to estimate the PSD [94, 116, 63], successful application of these methods has been demonstrated experimentally only for spheres [50] and octahedra [116]. Furthermore, it has been shown that the CLD inversion for high-aspect-ratio particles is highly ill-posed [116]. Finally, laser backscattering cannot measure the shape of individual particles and therefore cannot measure the distribution of particles between different shape classes (e.g. number of needles relative to number of spheres) nor shape factor distributions (e.g. distribution of aspect ratios).

The limitations inherent in laser-scattering-based monitoring technologies motivate the use of imaging-based methods, which allow direct visualization of particle size and shape. Obtaining quantitative information from imaging-based methods, however, requires image segmentation. Image segmentation means separating the objects of interest (e.g. the particles) from the background. Most commercial, imaging-based, on-line particle size and shape analyzers solve the segmentation problem by imaging the particulate slurry as it passes through a specially-designed flow cell under controlled lighting conditions [3, p. 167]. The images acquired in this way can be segmented using simple thresholding methods. The drawback is that this approach requires sampling, which is inconvenient, possibly hazardous, and raises concerns about whether the sample is representative of the bulk slurry. Alternatively, the images can be acquired in situ. Robust and efficient segmentation of in situ images is challenging, however, for several reasons. First, in situ imaging typically requires illumination by reflected light in order to handle high solids concentrations. Thus, the particles appear in the image with non-uniform color and intensity such that thresholding methods are ineffective. The use of reflected light can also result in poorly-defined particle outlines, thus limiting the robustness of methods based on closing the particle outlines, such as the technique proposed by [19]. Second, the crystals can have a large variation in size and

are randomly-oriented in 3-D space, which means the projections of the crystals onto the imaging plane can take on a wide variety of shapes. Hough-transform-based methods [41, p. 587], which have been applied extensively to segment images of circular or elliptical particles and droplets, involve exhaustive searches over the particle location and shape parameter space and are therefore computationally infeasible for randomly-oriented particles of complex shapes. The image segmentation problem is further complicated by particle agglomeration, overlap, breakage, and attrition, which result in occluded or poorly-defined particle boundaries.

Model-based object recognition is a widely-used approach to computer vision that has been developed to enable automatic recognition of complex objects with unknown pose (i.e. orientation with respect to the camera) in the presence of missing or occluded data [51, 65, 37, chapter 18]. The model-based object recognition approach is based on matching raw image features (such as arcs or lines) with one of several pre-defined models. The model-based approach does not involve exhaustive searches of the model parameter space and is therefore more efficient than purely top-down approaches (such as Hough-transform-based methods). Model-based object recognition is more robust to noise than purely bottom-up approaches because it can be applied even if part of the object to be identified is occluded or missing. Furthermore, the model-based approach leads to algorithms that can be implemented in a parallel fashion to enable real-time analysis. Algorithms based on this approach have been developed and applied to systems of circular particles [99], elliptical particles [47], and rectangular or needle-like particles [59].

This paper describes a model-based object recognition algorithm that can be applied to images of crystals of any shape, provided the shape can be represented as a wire-frame model. The wire-frame models used by the algorithm are parameterized. Thus, a single model can be used to identify crystal objects exhibiting a wide range of sizes and shapes within a given shape class. The algorithm therefore enables the measurement of shape factor distributions. Furthermore, the algorithm can be applied using multiple wire-frame models representing different shape classes to measure the distribution of particles between different shape classes. The algorithm's effectiveness is demonstrated by comparing the algorithm results with those obtained by manual, human analysis of in situ video images acquired at different solids concentrations during an α-glycine cooling crystallization experiment. The paper is organized as follows: Section 8.4 describes the crystallization experiment. Section 6.2 describes the algorithm, and Section 8.5 discusses the algorithm's accuracy and speed.

6.1 Experimental

The α-glycine crystallization was carried out by dissolving 180 g of glycine (p.a., Acros Organics) in 600 mL of deionized water at 55 °C. The solution was cooled to 25 °C at 5 °C/hr. Spontaneous nucleation was observed around 29 °C. Three sets of video images were acquired during the course of the experiment, each set consisting of 100 images acquired at a rate of 30 frames per second. Each set was acquired once a noticeable increase in the solids concentration had occurred. The first set of video images was acquired approximately 13 minutes after nucleation at low solids concentration with the draft tube in the middle of the crystallizer clearly visible. The second set

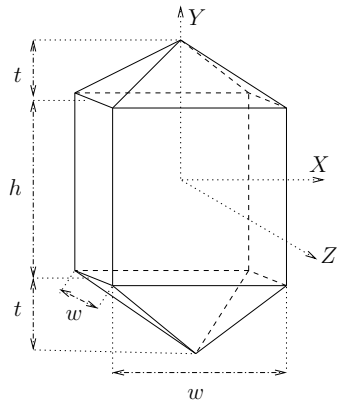


Figure S1: Wire-Frame glycine crystal model. The parameters for the model are the crystal body height, h, the width, w, and the pyramid height, t.

of images was acquired approximately 24 minutes after nucleation at medium solids concentration with the draft tube just barely visible. The final set of images was acquired 43 minutes after nucleation at high solids concentration with the draft tube completely obscured by the slurry. The polymorphic form of the observed crystals was confirmed using X-ray powder diffraction.

6.2 Model-based recognition algorithm

This section describes the model-based recognition algorithm designed to extract crystal size and shape information from in situ crystallization images. The algorithm is called M-SHARC (**M**odel-based **SHA**pe **R**ecognition for **C**rystals) and has been implemented in MATLAB 7.0.

6.2.1 Preliminaries

The model-based object recognition framework involves matching a set of primitive features extracted from an image (such as points, corners, or lines) to a pre-defined set of models. The primitive image features used by M-SHARC are lines, and the models are parameterized, wire-frame models. Wire-frame models consist of a set of q vertices $V = \{\mathbf{X}_K[\mathbf{p}_m]\}_{K=1\ldots q}$ and a set of r lines or edges $E = \{E_J\}_{J=1\ldots r}$. $\mathbf{X}_K$ is a three-dimensional vector defined in a model-centered coordinate system as a function of the model internal parameters $\mathbf{p}_m$. E_J is a set of two labels pointing to the vertices in V that are connected by edge J. The model used in this study was designed to capture the range of shapes exhibited by crystals of glycine, an amino acid of importance in the pharmaceutical industry. The model, shown in Figure S1, has three internal parameters ($\mathbf{p}_m = (h, w, t)$), 20 edges, and 10 vertices.

To fit the wire-frame model to the linear features in the image, the model must be projected onto the image plane. This projection is computed by first applying rigid-body rotations and translations to change each model point $\mathbf{X}$ from the model-centered coordinate frame to the

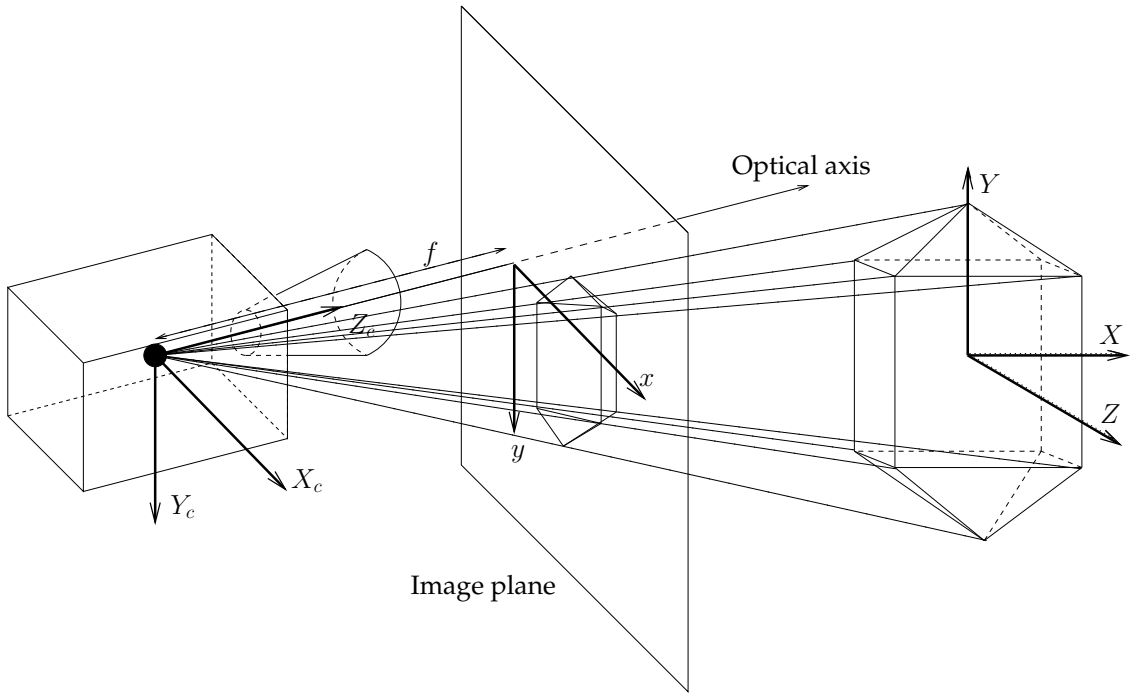


Figure S2: Depiction of the perspective projection of the glycine model onto the image plane. For simplicity, the image plane is displayed in front of the camera.

camera-centered coordinate frame:

$$\mathbf{X_c} = \mathbf{R_z R_y R_x X} + \mathbf{T} \tag{6.1}$$

in which $\mathbf{R_z}$, $\mathbf{R_y}$, and $\mathbf{R_x}$ are rigid-body rotation matrices, which are functions of the in-plane orientation θ_z and the orientations in depth θ_y and θ_x, respectively. $\mathbf{T} = (t_x, t_y, t_z)$ is a translation vector. Next, each model point is projected onto the image plane according to some imaging model. Under perspective projection, the transformation from a 3-D model point $\mathbf{X_c} = (X_c, Y_c, Z_c)$ to an image point $\mathbf{x} = (x, y)$ is given by

$$x = \frac{f}{Z_c} X_c, \quad y = \frac{f}{Z_c} Y_c \tag{6.2}$$

in which f is the focal length of the camera. Figure S2 depicts the perspective projection of the glycine model onto the image plane. M-SHARC uses the "weak perspective" imaging model, which accurately approximates perspective projection provided the depth of the imaged objects is small relative to the distance of the objects from the camera. Under such imaging conditions, f/Z_c and t_z can be assumed constant for all objects. In this work, we let $f/Z_c = 1$ and $t_z = 0$.

The projection of the model onto the image plane is completed by determining the lines that are visible for the given pose. Given a convex model, the visible model lines can be determined by computing an outward normal vector (in camera-centered coordinates) for each surface of the 3-D model. The sign of the dot product of this normal vector with the camera's optical axis determines

whether or not the surface is visible. The visible model lines are the lines that bound the visible surfaces.

The projection of the wire-frame model onto the image plane results in a set of projected model lines $E^P = \{(\mathbf{M}_J, \mathbf{T}_J, L_J)\}_{J=1...m}$ in which $\mathbf{M}_J$ is a vector pointing from the origin of the image coordinate system to the midpoint of the Jth model line, $\mathbf{T}_J$ is the unit tangent of the line, L_J is the length of the line, and m is the number of visible model lines. The set of data lines are defined similarly as $S = \{(\mathbf{m}_j, \mathbf{t}_j, l_j)\}_{j=1...n}$, in which n is the number of lines detected by the line finder.

M-SHARC follows the approach developed by [65, 66], consisting of three main steps: First, M-SHARC identifies linear features in the image. Second, M-SHARC identifies linear feature clusters that appear significant on the basis of viewpoint-independent relationships such as collinearity, parallelism, and end-point proximity. Third, M-SHARC fits a three-dimensional, wire-frame model to each significant linear feature cluster. The following sections describe each of these steps.

6.2.2 Linear feature detection

The M-SHARC algorithm uses the line finder proposed by [17], incorporating some of the speed-up suggestions given by [54] (see [59] for details). The Burns line finder detects lines by identifying regions of pixels having similar image intensity gradient orientation. By detecting lines on the basis of gradient orientation (as opposed to gradient magnitude), the Burns line finder's performance is relatively insensitive to variations in contrast and brightness. This property is important for crystallization imaging because, as crystallization occurs, the increasing solids concentration causes more reflected light to reach the camera CCD, which increases the image brightness while decreasing the image contrast.

During linear feature extraction, single physical edges are commonly broken up into multiple collinear lines due to particle overlap, noise, or poor lighting. A key component of the algorithm, therefore, is the grouping of these collinear lines prior to searching for viewpoint-invariant line groups. The M-SHARC algorithm uses the method developed by [31] because it is straightforward to implement and relatively inexpensive computationally. The method uses simple thresholding on the angle, perpendicular distance, and endpoint distance between two lines to determine whether or not the lines are collinear. If collinearity requirements are met, M-SHARC creates a new line based on the collinear pair but also retains the two original lines in case the instance of collinearity is accidental. Therefore, subsequent grouping operations operate on both the lines created from collinear grouping and all lines determined by the Burns line finder (whether or not they are involved in a collinear group). Retaining the original lines involved in collinear groups makes M-SHARC's performance less sensitive to the collinear grouping thresholds.

6.2.3 Perceptual grouping

Perceptual grouping refers to the task of organizing primitive objects, such as points or lines, into higher-level, meaningful structures. These structures, or groups, are useful as visual cues for the location, size, and orientation of a given object in the image. Viewpoint-invariant groups (VIGs),

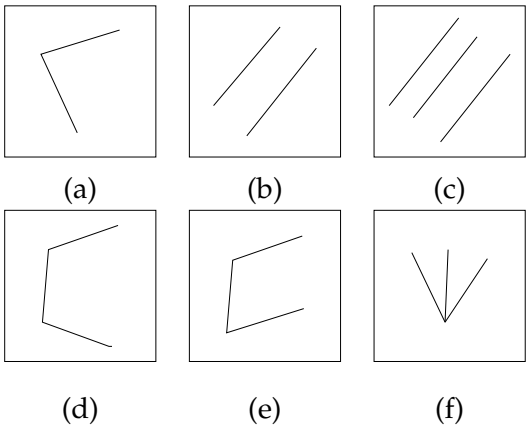


Figure S3: Depiction of different viewpoint-invariant line groups (VIGs) used by M-SHARC. (a) Junction. (b) Parallel pair. (c) Parallel triple. (d) C-triple. (e) C-square. (f) Arrow.

or groups that maintain certain properties regardless of the camera viewpoint, are necessary because the orientation of the object with respect to the camera is generally unknown. M-SHARC identifies line groups that can be assigned to one of the classifications depicted in Figure S3. The groups are identified based on orientation differences, spatial differences, and connectivities between the lines in the image. For example, junctions are line pairs that satisfy angle dissimilarity and endpoint proximity thresholds. Parallel line groups satisfy angle similarity and perpendicular distance thresholds. C-triples consist of three lines connected at two junctions where the angle between the lines at each junction is greater than 90 degrees. C-squares consist of three lines connected at two junctions with two of the lines being approximately parallel. Arrows are three lines connected at a single junction where the angles between lines are less than 90 degrees.

M-SHARC calculates a significance measure for each VIG based on the line lengths, endpoint distances, and (for groups of parallel lines) orientation differences, as described by [65]. This significance measure is used to ensure that the most visually salient VIGs are considered first in the model-fitting stage.

6.2.4 Model-fitting

The objective of the model-fitting process is to determine the model parameters and viewpoint parameters such that the two-dimensional projection of the geometric model matches the low-level features extracted from the image. The VIGs identified during the perceptual grouping stage provide the starting point for the tasks associated with the model-fitting process. The first task, called the correspondence problem, involves deciding which data lines correspond to which model lines. In general, multiple correspondences are possible, so M-SHARC may generate multiple hypotheses for a given VIG. Next, M-SHARC uses the data lines' positions and lengths to estimate the model and viewpoint parameters for each correspondence hypothesis. Given these parameters, M-SHARC projects the model into the image, identifies additional correspondences between model lines and data lines, and calculates a verification score for the set of correspon-

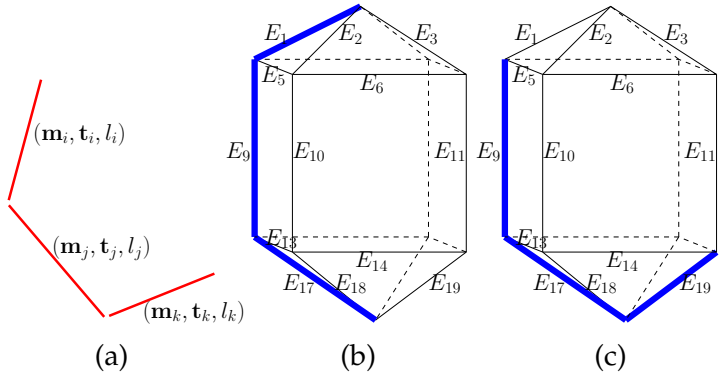


Figure S4: Depiction of two correspondence hypotheses. (a) Data line segments. (b) Hypothesis 1: Data lines i, j, and k correspond to model edges E_1, E_9, and E_{17}, respectively. (c) Hypothesis 2: Data lines i, j, and k correspond to model edges E_9, E_{17}, and E_{19}, respectively.

dences. In the case of multiple hypotheses, M-SHARC chooses the correspondence hypothesis that results in the highest verification score and performs a least-squares minimization to achieve a better fit between the model and data lines. Finally, M-SHARC invalidates any VIGs that contain lines that are completely enclosed within the bounding box of the projection of the optimized model. These tasks are described in the following subsections.

Determining correspondences

The first step in the model-fitting process is to determine which of the model lines from the set E correspond to the data lines in a given VIG. M-SHARC solves the correspondence problem by rotating the data lines in the VIG into a standard position such that unambiguous descriptors can be assigned to each line in the group. These descriptors are used to hypothesize one-to-one correspondences between each data line in the VIG and a model line. M-SHARC currently has methods for hypothesizing correspondences for parallel pairs, parallel triples, and C-triples. For example, a triple of parallel lines is rotated such that all lines are basically vertical, and each line is labeled as left-most, center, or right-most. These labels are used to assign each of the parallel data lines to their corresponding model lines. In the case of a C-triple, the lines are rotated such that the center line is vertical, and the lines are labeled as top, center, or bottom. The correspondence remains ambiguous, however, so multiple hypotheses must be tested, as depicted in Figure S4. In this figure, data lines i, j, and k could correspond, respectively, to either E_1, E_9, and E_{17}, or E_9, E_{17}, and E_{19}.

Estimating model and viewpoint parameters

The problem of estimating viewpoint parameters, also referred to as pose estimation or alignment, has been studied extensively in the literature. The research has focused on estimating pose from both point correspondences [48, 52] and line correspondences [26, 20, 115]. A major drawback

of these methods is that the internal model parameters are assumed known. The utility of these methods is therefore limited for the parameterized models of interest in our study.

M-SHARC estimates the internal model parameters using the properties of the data lines and an assumed orientation in depth. For example, given the first hypothesis in Figure S4, the model height h is estimated as $l_j \cos\theta_x$ and the width w is estimated as $2\cos\theta_y(\max(|\mathbf{t}_i \cdot \mathbf{t}_j^{\perp}|, |\mathbf{t}_k \cdot \mathbf{t}_j^{\perp}|)$, in which θ_y and θ_x are assumed orientations in depth and $\mathbf{t}_j^{\perp}$ is a unit vector perpendicular to $\mathbf{t}_j$. The pyramid height t is estimated as $w \tan\alpha/2$, in which α is assumed to be 45 degrees. M-SHARC currently has methods for estimating model parameters on the basis of correspondences for parallel pairs, parallel triples, and C-triples.

Given the correspondences, assumed orientations in depth, and model parameters, the remaining viewpoint parameters are estimated in a straightforward manner. First, M-SHARC projects the 3-D model onto the image plane using Equations (8.9) and (8.10), assuming $\theta_z = t_x = t_y = 0$. Next, the in-plane orientation θ_z is estimated using a weighted average of the orientation differences between the data lines and their corresponding, projected model lines. Given the estimated θ_z, the 3-D model is then projected once again to the image plane and the translations t_x and t_y are estimated from a weighted average of the spatial differences between the model line midpoints and their corresponding data line midpoints.

Identifying additional correspondences

As described in section 6.2.1, the projection of the wire-frame model onto the image plane results in a set of projected model lines E^P that can be compared directly with the set of data lines S. M-SHARC identifies additional model-data line correspondences by identifying instances of parallelism between the model lines and data lines.

Calculating the verification score

The purpose of the verification score is to quantify the amount of evidence in the image supporting a given correspondence hypothesis. M-SHARC calculates the verification score as the percentage of visible model line length that is supported or overlapped by data lines. The verification score threshold is set based on experimentation. A more advanced method for setting this threshold is described by [43], but their method is not generalized to parameterized models.

Minimizing model-data mismatch

Once a suitable hypothesis is identified, M-SHARC minimizes the mismatch between corresponding model and data lines by solving the optimization problem suggested by [66]

$$\min_{\mathbf{p}} \Phi = \sum_{i \in M} l_i (e_{1i} + e_{2i})^2 \tag{6.3}$$

subject to the model and imaging constraints (Equations (8.9) and (8.10)). In Equation (6.3), M is the set of data lines for which a correspondence has been identified, $\mathbf{p}$ is the parameter vector,

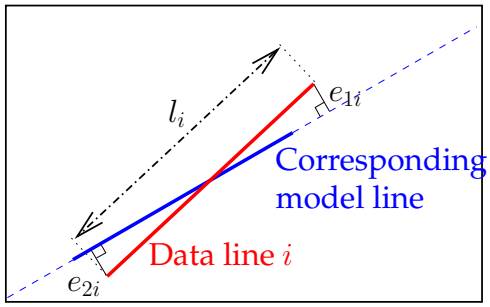


Figure S5: Depiction of variables used in mismatch calculation for a single line correspondence.

l_i is the length of data line i, and e_{1i} and e_{2i} are the perpendicular distances from data line i's endpoints to the corresponding model line, as depicted in Figure S5. The parameter vector for the glycine model in Figure S1 is

$$\mathbf{p} = \begin{bmatrix} h & t & w & \theta_x & \theta_y & \theta_z & t_x & t_y \end{bmatrix} \tag{6.4}$$

Invalidating overlapping VIGs

The VIGs identified during M-SHARC's perceptual grouping stage can overlap, sharing lines. We assume, however, that each line in the image can be attributed to only one crystal. Thus, once a model is successfully fit to a set of lines, any VIG that contains one or more of those lines is considered invalid. Furthermore, we assume that any line completely contained within the bounding box of an identified crystal has arisen due to unmodeled features of that crystal. Thus, any VIG that contains one or more of these lines is also considered invalid. M-SHARC does not attempt to fit a model to invalid VIGs.

Applying this constraint on overlapping VIGs is advantageous in that it significantly reduces the number of VIGs investigated, thus increasing M-SHARC's efficiency. Furthermore, initial studies indicated that a large number of false positives are identified if VIG invalidation is not employed. Thus, the overlapping crystal constraint serves to increase the algorithm's accuracy as well as reduce the computational burden such that the algorithm can run in real-time on a single processor. However, this feature requires M-SHARC to investigate each VIG serially, starting with the most significant VIG and proceeding to the least significant, such that M-SHARC currently cannot be implemented in parallel.

6.2.5 Summary and example

The M-SHARC algorithm can be summarized using the following pseudo-code:

```
1. Detect linear features in image
2. Identify and sort viewpoint-invariant line groups (VIG)
For each VIG
    If VIG is valid
```

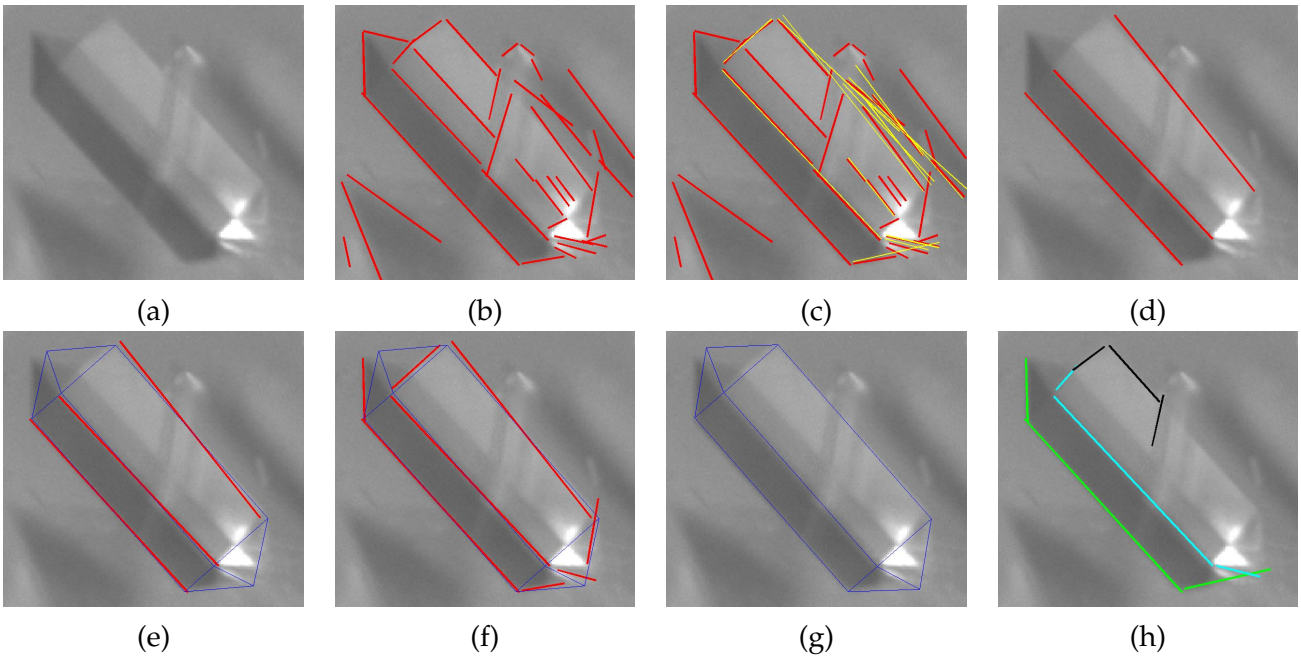


Figure S6: Result of applying M-SHARC to image of α-glycine crystal. (a) Original region of interest. (b) Linear features extracted using Burns line finder (dark lines). (c) Linear features extracted using collinear grouping (light lines). (d) Most salient line group. (e) Model initialization. (f) Identification of additional correspondences. (g) Optimized model fit. (h) Invalidated VIGs.

```
        3. Generate correspondence hypotheses
        For each hypothesis
            4. Estimate model and viewpoint parameters
            5. Project model into image
            6. Search for additional correspondences
            7. Compute verification score
        Endfor
        8. Select hypothesis with highest score
        If (score > Verification threshold)
            9. Minimize model-data mismatch
           10. Save optimized model information
           11. Invalidate overlapped VIGs
        Endif
    Endif
Endfor
```

Figure S6 shows the step-by-step results of applying the M-SHARC algorithm to a region of interest from an image of α-glycine crystals. This region of interest contains a large, well-defined crystal that is partially overlapped by a small, poorly-defined crystal. Applying the line finder to

the image in Figure S6(a) produces the dark lines shown in Figure S6(b). Because of overlap by the smaller crystal, a few of the edges of the larger crystal are broken up into two or more lines. Using collinear line grouping, these broken-up lines are combined into the single, light lines displayed in Figure S6(c). In the perceptual grouping stage, the line group that is identified as being most significant is the triple of parallel lines shown in Figure S6(d). The lengths, spatial positions, and relative distances between these lines provide the information necessary to estimate the internal model parameters and pose, resulting in the model projection shown in Figure S6(e). In this figure, the thin, solid lines are the visible model lines while the dashed lines are the non-visible model lines. Figure S6(f) shows the additional correspondences between model and data lines found by identifying instances of parallelism between the visible model lines in Figure S6(e) and the data lines in Figure S6(c). Figure S6(g) shows the optimized model obtained by minimizing the perpendicular distances between each of the model-data line correspondences. Finally Figure S6(h) indicates three of the VIGs that are invalidated due to overlap with the identified model.

6.3 Results

To assess our algorithm's performance, we carried out an unseeded, α-glycine cooling crystallization. Following spontaneous nucleation, we acquired three sets of video images at varying solids concentrations with 100 images in each set. Given the rate of acquisition (30 frames per second), each set of images essentially represents a snapshot in time. All images were analyzed by M-SHARC using the same set of parameters. The VIGs used to initialize model-fitting were parallel triples and C-triples.

6.3.1 Visual evaluation

Figures S7–S9 show the results of applying M-SHARC to images acquired at low, medium, and high solids concentrations. These figures demonstrate M-SHARC's ability to identify crystals covering a wide range of sizes and orientations. These figures also demonstrate M-SHARC's ability to handle poor image quality. Although several options exist for improving the images, such as lowering the mixing rate to decrease motion blur, successful image analysis of these low-quality images demonstrates M-SHARC's robustness and helps ensure success in the rugged industrial environment.

6.3.2 Comparison with human analysis

To quantify M-SHARC's performance, we compare its results with the results obtained by manual analysis of the images by human operators. All 300 images acquired during the α-glycine crystallization were analyzed both by M-SHARC and by a human operator. The human operator annotated the images using LabelMe, a database and web-based image annotation tool developed by [95]. The images and annotations are available to the general scientific community through the LabelMe website.

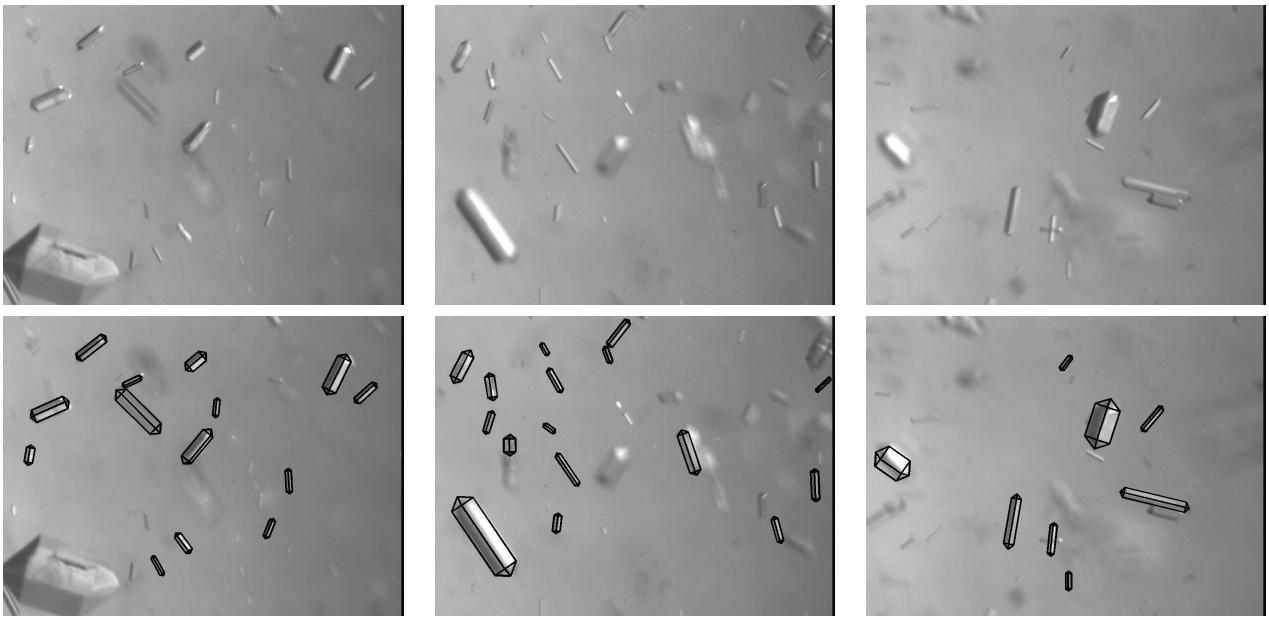


Figure S7: M-SHARC segmentation results for selected images acquired at low solids concentration (13 min. after appearance of crystals).

Figure S10 illustrates how M-SHARC's results are compared with the human operator's results. Figure S10(b) shows the crystal outlines determined by the human operator while Figure S10(c) shows the outlines determined by M-SHARC. Each outline is simply a set of straight lines. If a sufficiently large set of correspondences can be found between the set of lines in an M-SHARC outline and the set of lines in a human outline, the crystal corresponding to those outlines is classified as a *hit*. If a sufficiently large set of correspondences is *not* found between a given M-SHARC outline and any other human outline, the crystal corresponding to the M-SHARC outline is classified as a *false positive*. Similarly, a crystal identified by a human operator for which no corresponding M-SHARC outline can be found is classified as a *miss*. In Figure S10(d), the false positives are displayed in white and the misses in black.

Table 6.1 shows the number of hits N_H, misses N_M, and false positives N_{FP} identified by comparing M-SHARC's results with the human operator's results for each of the three sets of video images, along with number and area fractions for the hits and false positives. The hit, miss, and false positive areas (A_H, A_M, and A_{FP}, respectively) are calculated based on the areas of the polygons defined by each crystal outline. For each level of solids concentration, the hit number fraction is significantly less than the hit area fraction while the false positive number fraction is comparable to the false positive area fraction. Thus, we would expect the number-based area distribution to be biased towards larger particles. This expectation is verified by Figure S11, which compares the number-based cumulative distribution functions (CDFs) for particle area. As a standard of comparison, Figure S11 also shows the CDFs constructed using only particles classified as hits. The confidence interval displayed in these figures is calculated using the Kolmogorov-

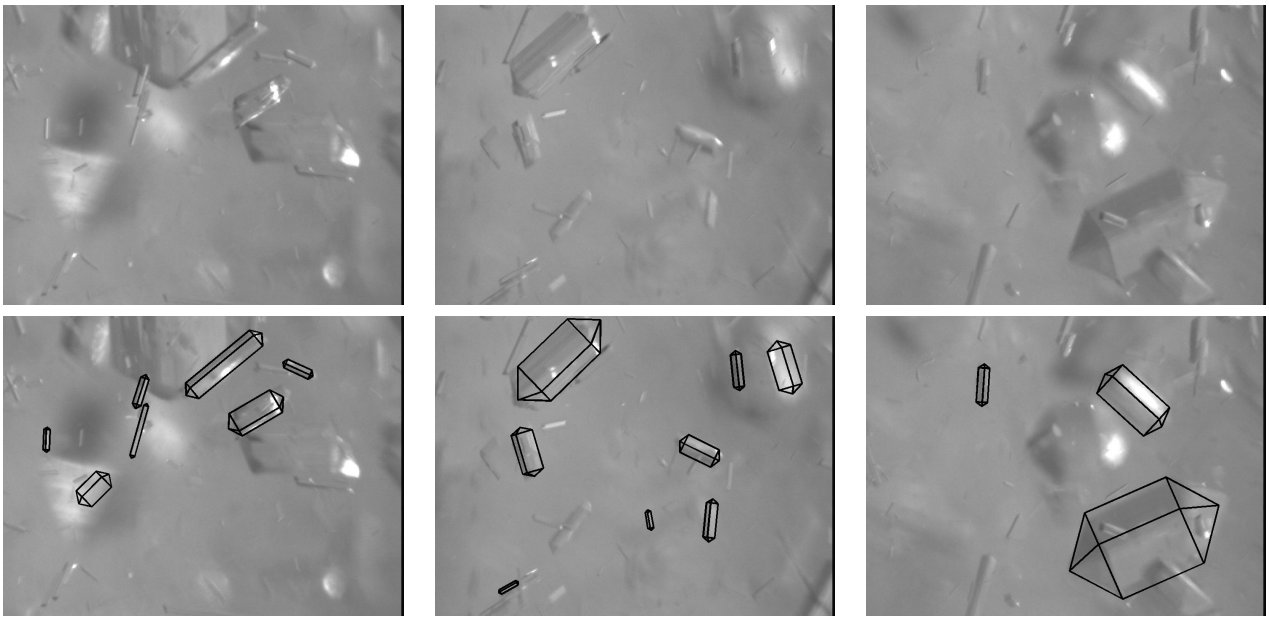


Figure S8: M-SHARC segmentation results for selected images acquired at medium solids concentration (24 min. after appearance of crystals).

	Low	Med.	High
Hits (N_H)	500	279	220
Misses (N_M)	514	657	445
False Positives (N_{FP})	130	191	352
Hit number fraction ($N_H/(N_H + N_M)$)	0.49	0.30	0.33
False Pos. number fraction ($N_{FP}/(N_H + N_{FP})$)	0.21	0.41	0.62
Hit area fraction ($A_H/(A_H + A_M)$)	0.63	0.38	0.31
False Pos. area fraction ($A_{FP}/(A_H + A_{FP})$)	0.23	0.36	0.53

Table 6.1: Summary of comparison between M-SHARC results and human operator results for in situ video images obtained at low, medium, and high solids concentrations (100 images at each concentration).

Smirnov statistic [71, 23].

To identify possible improvements to the M-SHARC algorithm, we examined the results of M-SHARC's subroutines (line finding, perceptual grouping, and model fitting) for the 100 largest misses and the 100 largest false positives. Examination of the large misses reveals that many of the misses are particles that are somewhat blurry with low contrast edges that are difficult for M-SHARC's line finder to identify (see row 1 in Figure S12). Despite the low contrast, the visual cues are sufficient that the outline of these particles can be discerned by a human operator. Particles that are clearly in-focus also sometimes exhibit low contrast edges and cause difficulties for M-SHARC's line finder (see row 2 in Figure S12). Particle agglomeration is another major source

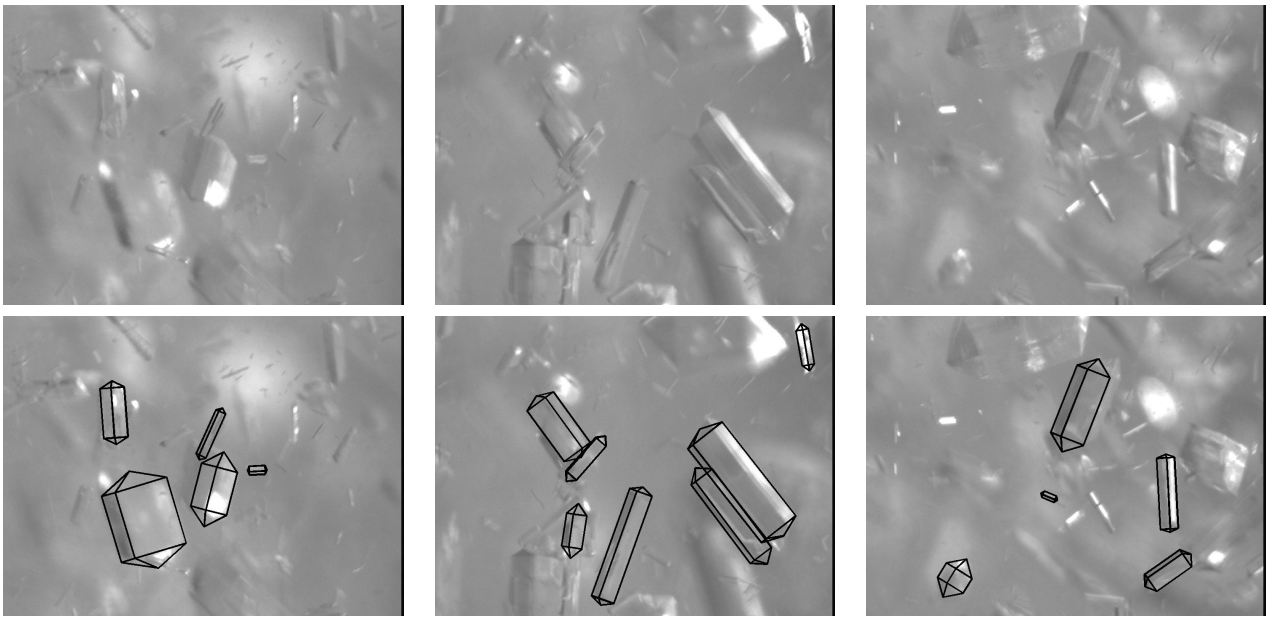
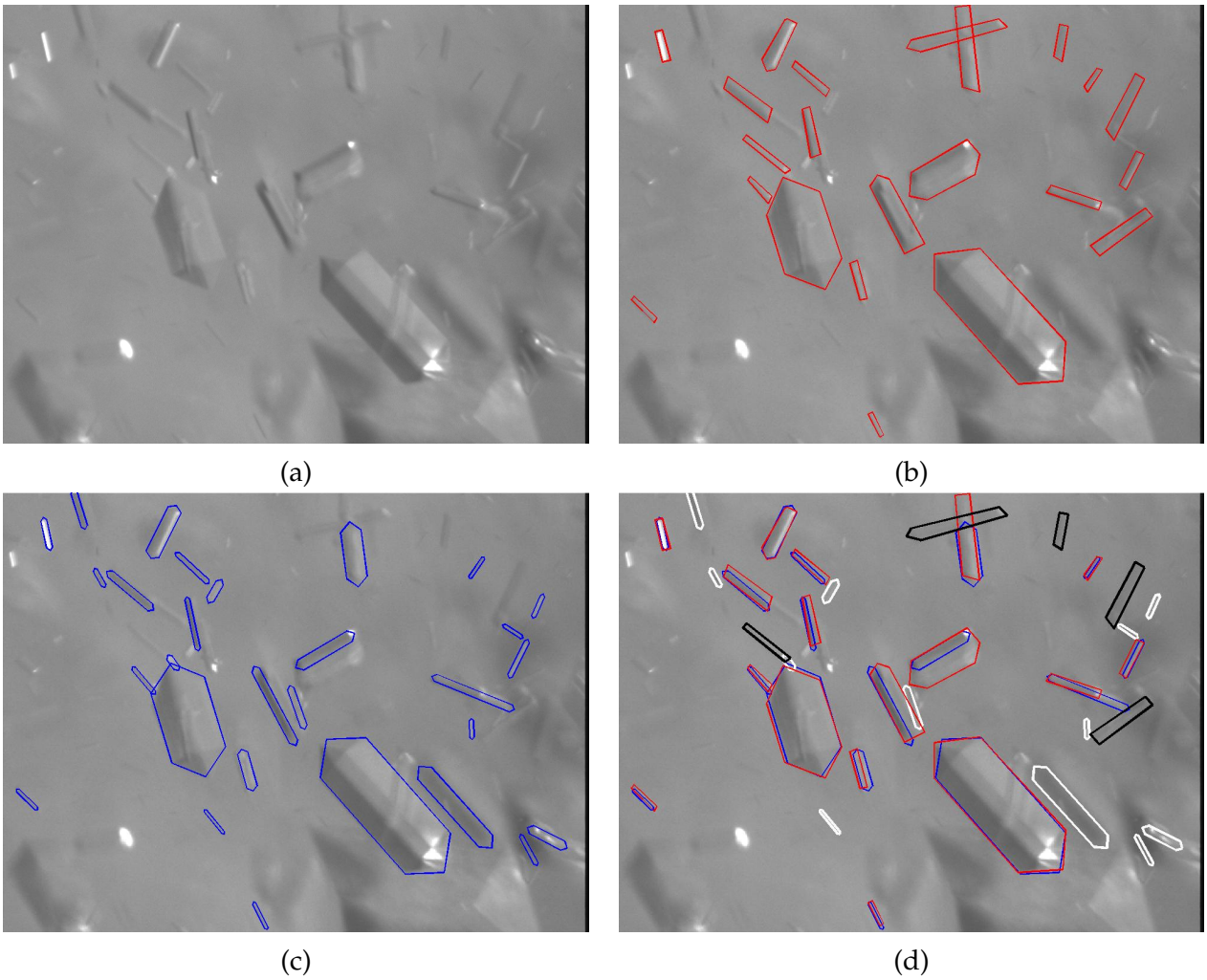


Figure S9: M-SHARC segmentation results for selected images acquired at high solids concentration (43 min. after appearance of crystals).

of difficulty for M-SHARC, causing the particle edges to be broken up at each point of agglomeration. For agglomerated particles, the line-finding routine results in a concentrated group of small linear features, complicating the perceptual grouping stage (see row 3 in Figure S12). The degree of failure at the line-finding stage due to blur and agglomeration usually makes successful perceptual grouping an unreasonable goal. In some cases, salient VIGs can be found but have not been utilized because routines for solving the correspondence and model parameter estimation problems have not yet been developed. Row 4 in Figure S12 gives two examples of crystals resulting in significant VIGs for which initialization routines have yet to be developed. The development of such routines may be the best approach to minimize the number of misses.

Examination of the 100 largest false positives showed that a large fraction (more than 1/3) of the false positives are agglomerated or blurry crystals. The ambiguity associated with determining an outline for blurry and agglomerated crystals caused the human operators to pass over these crystals. Whether or not these false positives should be counted as false positives is unclear. The other major source of false positives arises due to shortcomings in M-SHARC's method of verification. For many of the false positives, the model and data lines do not align well, indicating that the orientational and spatial offset thresholds used in the verification process are not sufficiently stringent. Another indication that more stringent thresholds are necessary in the verification stage is that many of the false positives have only 3 or 4 data lines that correspond to the model lines. Unfortunately, this is also true for many of the hits, partly because the wire-frame model is only a rough approximation of the α-glycine shape. Perhaps using wire-frame model that more accurately represents the α-glycine shape would result in better fits for the hits, enabling the



(a) (b) (c) (d)

Figure S10: Illustration of comparison between human operator results and M-SHARC results. (a) Original image. (b) Outlines of crystals identified by human operator. (c) Outlines of crystals identified by M-SHARC. (d) Result of comparison between human outlines and M-SHARC outlines. Crystals identified as false positives are outlined in white while those identified as misses are outlined in black.

use of more stringent verification thresholds to eliminate false positives.

6.3.3 Algorithm speed

The average cputimes required to analyze images from each of the three video sets from the α-glycine experiment are shown in Table 6.2. The amount of time between each video acquisition during the crystallization experiment was approximately 12 minutes. Based on the cputimes in Table 6.2, M-SHARC can analyze approximately 10 images per minute, or 100 images every 10 minutes. This analysis speed is sufficiently fast for real-time implementation on chemical systems

Set	Line Finding	Group Finding	Initialization	Optimization	Total
1	2.6 (55)	1.2 (25)	0.2 (5)	0.7 (14)	4.8
2	3.3 (43)	3.3 (43)	0.3 (4)	0.7 (9)	7.6
3	3.3 (16)	15.4 (76)	0.4 (2)	1.0 (5)	20.2

Table 6.2: Average cputime required to analyze single image for three different image sets of increasing solids concentration. The first number in each number pair is the cputime in seconds. The second, parenthesized number gives the percentage of total cputime.

with crystallization dynamics similar to glycine.

The results presented in this paper are based on using a single wire-frame model. M-SHARC can be applied, however, using multiple models to represent different shape classes. Using multiple models would affect the computational requirements in two ways. First, each additional model may require the identification of additional VIGs to serve as visual cues for that model. However, the increase in computation time for the VIG identification stage may be small because many VIGs are identified using the same set of logical decisions. For instance, to identify a C-triple, M-SHARC must determine whether a given triple of lines is connected at one or two junctions. This same determination must be made to identify a C-square and an arrow. Second, the number of hypotheses would likely increase linearly with the number of models, resulting in a linear increase of computational requirements for the model initialization stage. In Table 6.2, the requirements for the initialization stage represent a small fraction of the total computational requirements, indicating that M-SHARC could accommodate several more models and remain suitably fast for real-time implementation.

6.4 Conclusions

We have developed a model-based object recognition algorithm that is effective for extracting crystal size and shape information from noisy, in situ crystallization images. The algorithm's accuracy has been assessed by comparing its results with those obtained by manual, human analysis of the images. At low solids concentrations, the algorithm identifies approximately half of the crystals identified by humans, while at medium to high solids, the algorithm identifies approximately one-third of the crystals. At low solids, false positives constitute approximately one-fifth of all identified crystals, while at medium to high concentrations the false positives constitute approximately half of the identified crystals. Despite the misses and false positives, the algorithm's cumulative size distribution measurements compare favorably with measurements obtained by humans but are biased towards larger particles. To improve the algorithm's accuracy, further development should focus on the algorithm's verification stage and on creating initialization routines for additional viewpoint-invariant line groups. The algorithm is sufficiently fast to provide on-line measurements for typical cooling crystallization processes.

6.5 Acknowledgment

This material is based upon work supported by the National Science Foundation under Grant No. 0540147. The authors gratefully acknowledge GlaxoSmithKline for providing the imaging system used in this study and Dr. Karen Jones of the University of Wisconsin School of Pharmacy for XRPD analysis of the glycine crystals. The authors also express appreciation to Dr. Antonio Torralba of the Massachussetts Institute of Technology Computer Science and Artificial Intelligence Laboratory for enabling the use of the LabelMe database and Susan Larsen for her help in manually tracing the crystals.

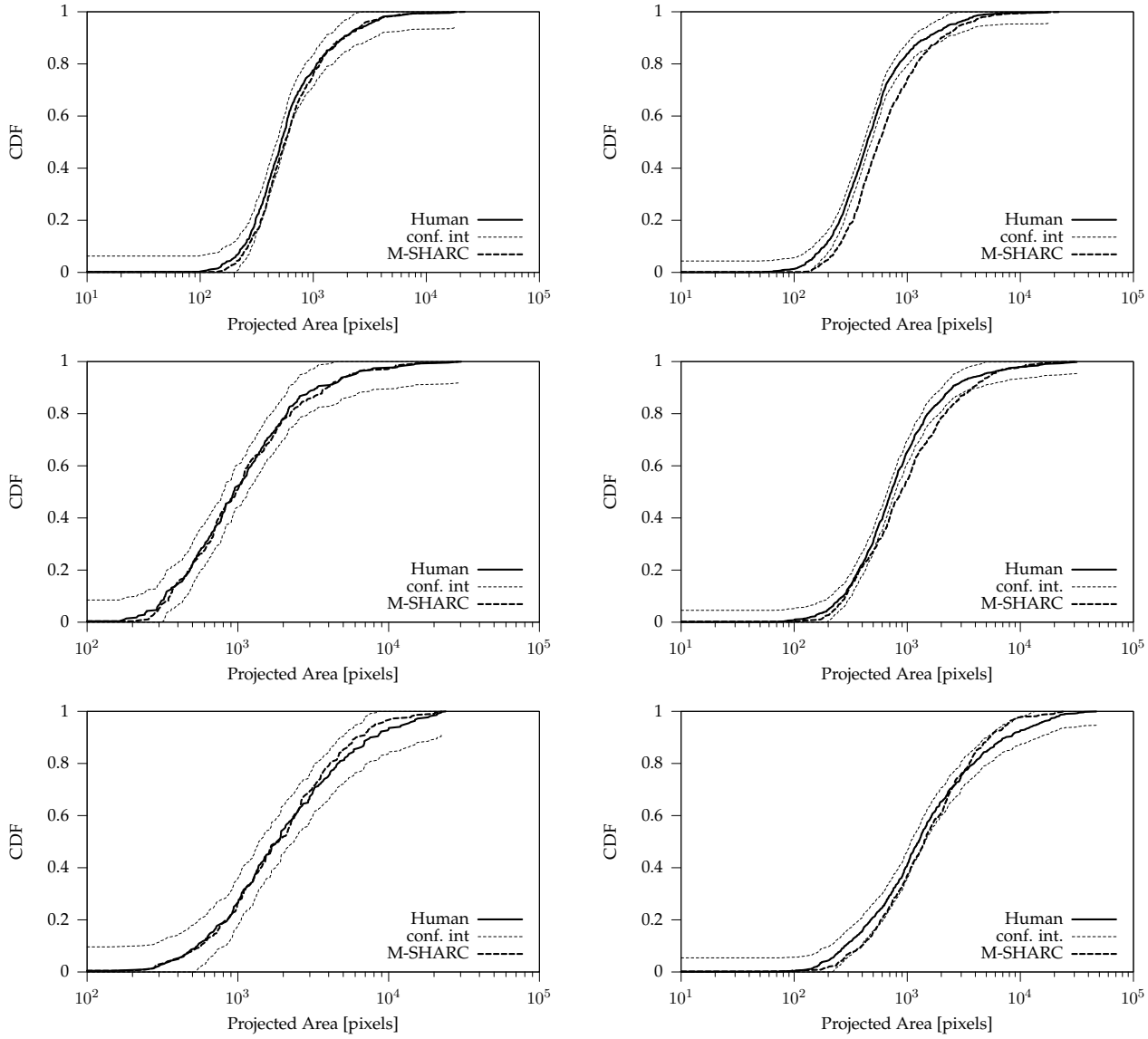


Figure S11: Comparison of Human and M-SHARC cumulative distribution functions for projected area. Rows 1, 2, and 3 show results for the α-glycine experiment at low, medium, and high solids concentrations, respectively. Column 1: CDFs constructed using only crystals classified as hits. Column 2: CDFs constructed using all crystals (i.e. the Human CDF is based on crystals classified as either hit or miss, while the M-SHARC CDF is based on crystals classified as either hit or false positive.

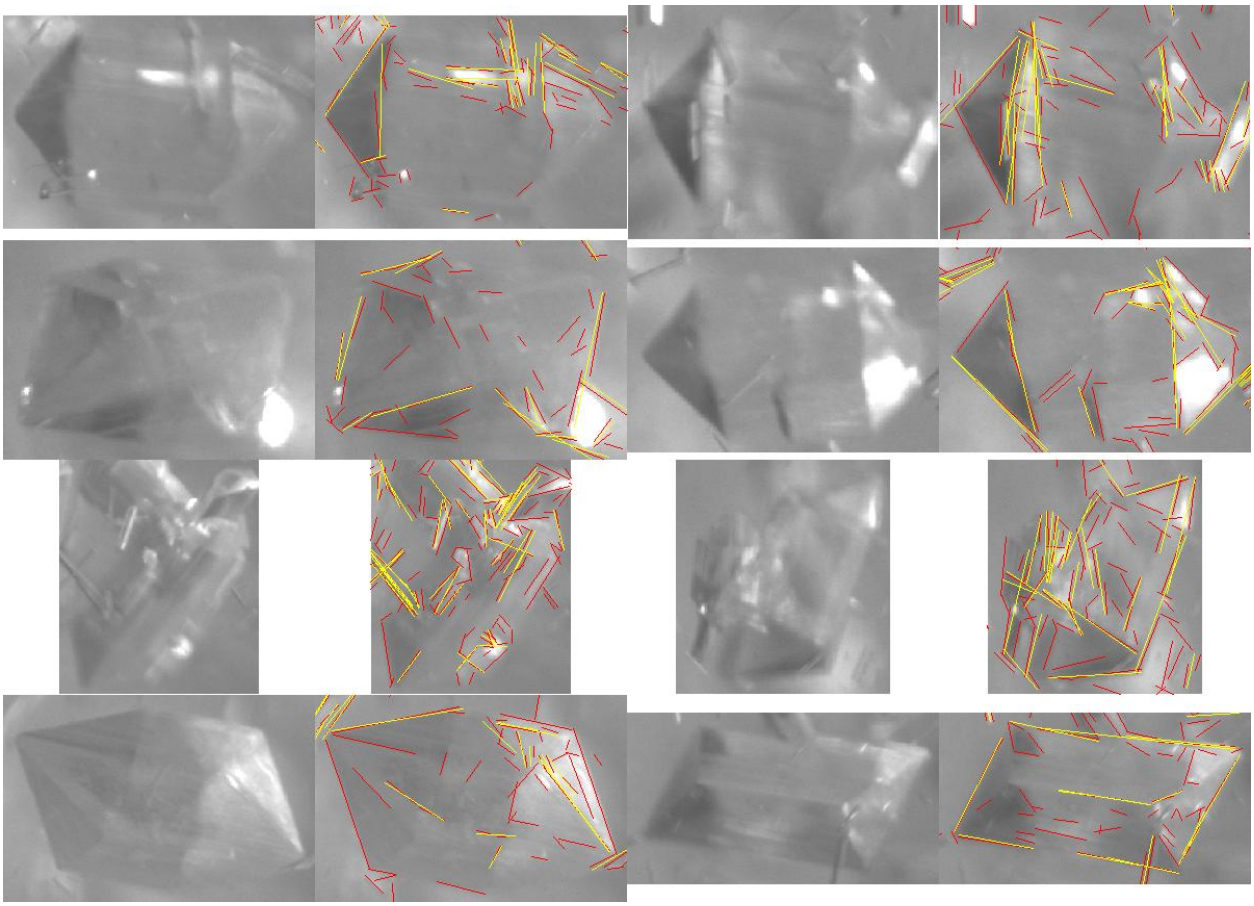


Figure S12: Results of linear feature detection for selected crystals missed by M-SHARC. The poor contrast for the crystals in row 1 is due to out-of-focus blur. The crystals in row 2 also exhibit poor contrast despite being seemingly in-focus. The crystals in row 3 show examples of agglomeration. The crystals in row 4 may be identifiable given further development of M-SHARC's correspondence and model parameter estimation routines described in Sections 6.2.4 and 6.2.4.

Chapter 7

Statistical estimation of PSD from imaging data

7.1 Introduction

Particulate processes are ubiquitous in the chemical, food, and pharmaceutical industries. A fundamental task for understanding and controlling particulate processes is measuring the particle size distribution (PSD). Historically, this task has been accomplished by mechanical methods such as sieving, but in the past several decades a wide variety of alternative technologies have been developed to make PSD measurement more efficient and more accurate. These technologies include, for example, forward laser light scattering and electrozone sensing. Unfortunately, these methods require sampling the particulate slurry, which is inconvenient and, in some cases, hazardous. Furthermore, obtaining a representative sample can be non-trivial. The difficulties associated with sampling have motivated the development of in situ sensors to measure PSD. These sensors include laser backscattering, ultrasonic spectroscopy, and in situ microscopy. Of these in situ sensors, only microscopy allows direct observation of particle shape, which makes this method advantageous for systems in which the particles are non-spherical.

Two challenges must be addressed to use in situ microscopy for PSD measurement. The first challenge is achieving image *segmentation*, or separating the objects of interest (e.g. the particles) from the background. Methods for segmenting in situ images have been developed for circular particles [99], arbitrarily-shaped particles [19], elliptical particles [47], needle-like particles [59], and particles with shape that can be represented using a wireframe model [60].

The second challenge is to estimate the PSD given the size and shape information obtained through successful image segmentation. Each segmented particle provides a single observation, which can be either *censored* or *uncensored*. A censored observation refers to an observation in which only partial information is obtained. For example, an observation of a particle touching the image border is censored because only a portion of the particle is visible. An observation of a particle with one end partially occluded by another particle is also censored. An observation is uncensored only if the particle is enclosed entirely within the image frame, is not occluded by other particles, and is oriented in a plane perpendicular to the optical axis of the camera.

A natural approach to estimate the PSD is to count only those particles appearing entirely within the field of view, not touching the image boundary. This approach, called minus-sampling, introduces sampling bias because the probability of counting a particle depends on its size and shape. For example, a small particle randomly located in the image has a high probability of

appearing entirely within the field of view, while a sufficiently large particle randomly located in the image may have zero probability of appearing entirely within the field of view. Miles [72] presented the first treatment of spatial sampling bias, developing a minus-sampling estimator that corrects spatial sampling bias by weighting each observation by M^{-1}, with M being related to the sampling probability of the observed particle. Miles derived formulas for M assuming a circular sampling region. Lantuejoul[57] extended Miles' results by showing how to calculate M for a rectangular sampling region.

The primary drawback of the Miles-Lantuejoul approach is that it uses only uncensored observations. If the size of the particles is large relative to the size of the image window, using censored observations (i.e. particles touching the image border) would be expected to result in improved PSD estimation. The primary goal of this study is to develop a PSD estimator using both censored and uncensored observations and to evaluate the benefits and drawbacks of this estimator relative to the Miles-Lantuejoul approach. We assume the censoring is due only to particles touching the image border and not due to orientation or occlusion effects. A secondary goal is to develop practical methods for determining confidence intervals for the estimated PSD. The methods developed in this study are intended for systems of high-aspect-ratio particles, which are commonplace in the pharmaceutical and specialty chemical industries.

The paper is organized as follows. Section 8.2 describes previous work related to PSD estimation of high-aspect-ratio particles and describes the application of the Miles-Lantuejoul estimator. Section 8.3 presents the formulation of the maximum likelihood PSD estimator and Section 8.4 describes the simulation studies used to test the estimator. Section 8.5 presents the results of these simulation studies, and Section 8.6 summarizes our findings. The full derivation of the maximum likelihood estimator can be found in [58].

7.2 Previous work

PSD estimation for high-aspect-ratio particles using imaging-based measurements is related to the problem of estimating the cumulative length distribution function H of line segments observed through a window, which has been investigated by several researchers. Laslett [61] was the first to derive the log likelihood for this problem. Wijer [113] derived the non-parametric maximum likelihood estimator (NPMLE) of H for a circular sampling region and an unknown orientation distribution function K. For arbitrary convex sampling regions, Wijer shows how to estimate H assuming K is known. Van Der Laan[106] studies the NPMLE of H for the one-dimensional line segment problem (i.e. all line segments have same orientation) for a non-convex sampling window, and Van Zwet[107] derives the NPMLE of H for the two-dimensional problem with a non-convex, highly irregular sampling region and known K. Svensson et al.[104] derive an estimator for a parametric length density function h using one-dimensional line segment data from a circular sampling region. Hall [45] derived an estimator for the intensity of a planar Poisson line segment process that is unbiased for any convex sampling region and any length distribution function. All of the above studies utilize both censored and uncensored observations. Baddeley [5] provides an excellent review of various spatial sampling estimation studies.

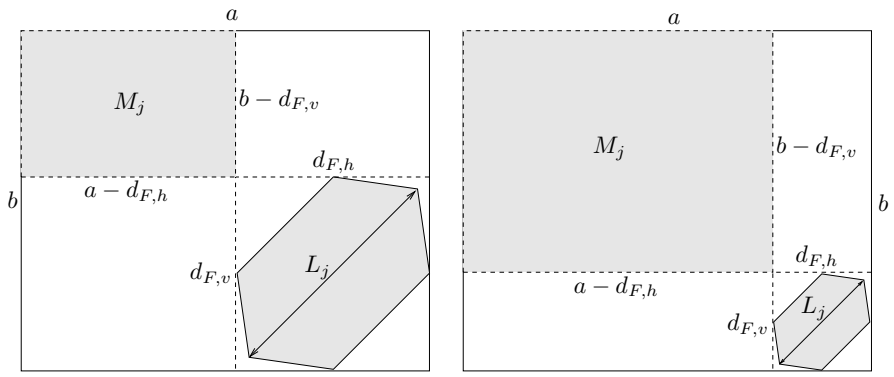


Figure S1: Depiction of methodology for calculating Miles-Lantuejoul M-values for particles of different lengths observed in an image of dimension $b \times a$.

The goal of the current study is to estimate the particle size distribution f, which is related to but different than the cumulative distribution function H or corresponding density function h for a line segment process. The PSD $f(L)$ is the number of particles of length L per unit volume and is related to H via the relation $H(L \leq l) = \int_0^l f(L)dL / \int_0^\infty f(L)dL$. The approach commonly used in practice to estimate the PSD from imaging-based measurements is the Miles-Lantuejoul method [72]. As there is some confusion amongst practitioners regarding the implementation of the Miles-Lantuejoul method, we describe the method here.

Let E^2 be the Euclidean plane, and let $D \subset E^2$ be a domain parameterized by (z, n, θ_n), in which z gives the center point of the domain, n gives the class, and θ_n is a vector giving the parameters necessary to completely specify a domain of class n. Let $Q(x) \subset E^2$ be a sampling region centered at x. For each domain D, define the set

$$D^\alpha = \{x \in E^2 : D \subset Q(x)\} \tag{7.1}$$

Thus, D^α is a domain comprising all points at which the sampling region can be placed and enclose entirely the domain D.

Define $M\{D\} = A\{D^\alpha\}$, where $A\{\cdot\}$ denotes area. Let $M\{D_j\}, j = 1 \ldots n$ be the M-values calculated for n observations of particles with lengths corresponding to size class i. Miles showed that $\rho_{ML_i} = \sum_j M\{D_j\}$ is an unbiased estimator of ρ_i, the density of particles in size class i per area, provided the minimum possible M-value for a particle in size class i is greater than zero.

In Miles' original paper [72], he derived formulas for calculating M-values for arbitrary domains assuming a circular sampling region. Later, Lantuejoul[57] extended Miles' results by showing how to calculate M for the rectangular sampling region typical of microscopy applications. For an image of size $a \times b$, the M-value of a particle is calculated by subtracting the vertical and horizontal Feret diameters of the particle ($d_{F,v}$ and $d_{F,h}$) from, respectively, the vertical and horizontal image dimensions b and a, as depicted in Figure S1.

7.3 Theory

7.3.1 PSD Definition

Consider a population of cylindrical or rod-like particles. The geometry of each particle is specified in terms of the cylinder height h and radius r. Define the characteristic length for the population of particles as $L = h$.

Consider a slurry of volume V in which a solid phase of discrete particles is dispersed in a continuous fluid phase. Let $f(L)$ denote the continuous PSD, or the number of particles of characteristic length L per unit volume slurry. Number-based PSDs are typically measured by discretizing the characteristic length scale into T non-overlapping bins or size classes. We therefore define the discrete PSD as

$$\rho_i = \int_{S_i}^{S_{i+1}} f(l)dl, \quad i = 1, \ldots, T \tag{7.2}$$

in which $\mathbf{S} = (S_1, \ldots, S_{T+1})$ is the vector of breaks between size classes.

The relative PSD $\mathbf{q}$ is a vector with elements

$$q_i = \frac{\rho_i}{\sum_j^T \rho_j}, i = 1, \ldots, T \tag{7.3}$$

In this paper, the term PSD is assumed to refer to the discrete, absolute PSD ρ unless specifically noted otherwise.

7.3.2 Sampling model

The particle population is sampled using in situ imaging. Let $V_I \in V$ denote the imaging volume, and assume V_I is a rectangular region of dimensions $a \times b \times d$, in which a is the horizontal image dimension, b is the vertical image dimension, and d is the depth of field. a and b determine the field of view, and we assume $a \geq b$. A single random sample of the population consists of an image containing the two-dimensional projection of the portions of particles inside V_I. We assume the system is well-mixed such that the centroids of the particles are randomly and uniformly distributed in space.

We assume the camera is positioned a fixed distance z_0 from the imaging volume, and that $d \ll z_0$. This assumption means the particles in the imaging volume are projected onto the image plane according to the weak perspective projection model. In other words, the projected particle lengths measured in the image coordinate system can be related to the true projected particle lengths by applying a constant magnification factor m.

We assume all particles are oriented in a plane orthogonal to the camera's optical axis. This assumption, together with the weak perspective assumption, essentially reduces the 3-D process to a 2-D process, thereby simplifying the analysis considerably. These assumptions are not used only for convenience, however, but rather to reflect the actual conditions under which in situ imaging measurements are made in practice. To obtain useful in situ images in high solids concentrations, the camera must have a small depth of field and be focused only a small depth into

the particulate slurry. It seems reasonable, therefore, to expect the shear flow at the slurry-sensor interface to cause the particles to align orthogonal to the this interface, and thus orthogonal to the camera's optical axis.

7.3.3 Maximum likelihood estimation of PSD

Let $\mathbf{X}_k = (X_{1k}, \ldots, X_{Tk})$ be a T-dimensional random vector in which X_{ik} gives the number of *non-border* particles of size class i observed in image k. A non-border particle is a particle that is completely enclosed within the imaging volume. A *border* particle, on the other hand, is only partially enclosed within the imaging volume such that only a portion of the particle is observable. For border particles, only the observed length (i.e. the length of the portion of the particle that is inside the imaging volume) can be measured. Accordingly, we let $\mathbf{Y}_k = (Y_{1k}, \ldots, Y_{Tk})$ be a T-dimensional random vector in which Y_{jk} gives the number of *border* particles with observed lengths in size class j that are observed in image k. We denote the observed data, or the realizations of the random vectors $\mathbf{X}_k$ and $\mathbf{Y}_k$, as $\mathbf{x}_k$ and $\mathbf{y}_k$, respectively.

The particle population is represented completely by the vectors $\rho = (\rho_i, \ldots, \rho_T)$ and $\mathbf{S} = (S_1, \ldots, S_{T+1})$ in which ρ_i represents the number of particles of size class i per unit volume and S_i is the lower bound of size class i. Given the data $\mathbf{x}$ and $\mathbf{y}$ (the subscript k denoting the image index is removed for simplicity), the maximum likelihood estimator of ρ is defined as

$$\rho_b = \arg\max_{\rho} p_{XY}(x_1, y_1, x_2, y_2, \ldots, x_T, y_T|\rho) \tag{7.4}$$

in which the subscript b indicates the use of border particle measurements and p_{XY} is the joint probability density for $\mathbf{X}$ and $\mathbf{Y}$. In other words, we want to determine the value of ρ that maximizes the probability of observing exactly x_1 non-border particles of size class 1, y_1 border particles of size class 1, x_2 non-border particles of size class 2, y_2 border particles of size class 2, and so on.

A simplified expression for p_{XY} can be obtained by noting that, at least at low solids concentrations, the observations X_1, Y_1, …, X_T, Y_T can be assumed to be independent. This assumption means that the observed number of particles of a given size class depends only on the density of particles in that same size class. At high solids concentrations, this assumption seems unreasonable because the number of particle observations in a given size class is reduced due to occlusions by particles in other size classes. At low concentrations, however, the likelihood of occlusion is low. The independence assumption does *not* imply that the observations are not correlated. Rather, the assumption implies that any correlation between observations is due to their dependence on a common set of parameters. As an example, if we observe a large number of non-border particles, we would expect to also observe a large number of border particles. This correlation can be explained by noting that the probability densities for both border and non-border observations depend on a common parameter, namely, the density of particles. Given the independence assumption, we express the likelihood function $L(\rho)$ as

$$L(\rho) = p_{XY} = \prod_{i=1}^{T} p_{X_i}(x_i|\rho) \prod_{j=1}^{T} p_{Y_j}(y_j|\rho) \tag{7.5}$$

in which p_{X_i} and p_{Y_j} are the probability densities for the random variables X_i and Y_j. The log likelihood is defined as $l(\rho) = \log L(\rho)$. Maximizing the likelihood function is equivalent to minimizing the log likelhood. Using Equation (A.2), the estimator in Equation (A.1) can therefore be reformulated as

$$\rho_b = \arg\min_{\rho} \sum_{i=1}^{T} -\log p_{X_i}(x_i|\rho) - \sum_{j=1}^{T} \log p_{Y_j}(y_j|\rho) \tag{7.6}$$

The probability densities p_{X_i} and p_{Y_j} can be derived given the particle geometry and the spatial and orientational probability distributions. In [58], p_{X_i} and p_{Y_j} are derived assuming the particles have needle-like geometry, are uniformly distributed in space, and are uniformly distributed in orientation. These derivations show that $X_i \sim \text{Poisson}(m_{X_i})$, or that X_i has a Poisson distribution with parameter $m_{X_i} = \rho_i \alpha_i$, in which α_i is a function of the field of view, depth of field, and the lower and upper bounds of size class i. Furthermore, $Y_j \sim \text{Poisson}(m_{Y_j})$, in which $m_{Y_j} = \sum_{i=1}^{T} \rho_i \beta_{ij}$

To extend the analysis to data collected from N images, we define two new random vectors $\mathbf{X_\Sigma}$ and $\mathbf{Y_\Sigma}$ for which $X_{\Sigma_i} = \sum_{k=1}^{N} X_{ik}$ and $Y_{\Sigma_j} = \sum_{k=1}^{N} Y_{jk}$. Here, the subscript k denotes the image index. Given that $X_{ik} \sim \text{Poisson}(m_{X_i})$, it can be shown that $X_{\Sigma_i} \sim \text{Poisson}(N m_{X_i})$ [12, p. 440]. Likewise, $Y_{\Sigma_j} \sim \text{Poisson}(N m_{Y_j})$.

Differentiating Equation (A.3) with respect to ρ and equating with zero results in a set of coupled, nonlinear equations for which an analytical solution is not apparent. Equation (A.3) is solved using MATLAB's nonlinear optimization solver FMINCON with initial values obtained from Equation (7.7).

If the border particles are ignored, the estimator reduces to

$$\rho = \arg\min_{\rho} \sum_{i=1}^{T} \left(-\log p_{X_i}(x_i|\rho)\right)$$

In this case, we can solve for ρ analytically:

$$\rho_i = \frac{X_i}{\alpha_i}, \quad i = 1, \ldots, T \tag{7.7}$$

The probability density for this estimator can be computed analytically as

$$p_{\rho_i}(\rho_i) = p_{\rho_i}(X_i/\alpha_i) = p_{X_i}(x_i) \tag{7.8}$$

with x_i being a non-negative integer. It is straightforward to show that this estimator has the following properties:

$$\begin{aligned} E[\rho_i] &= \rho_i \\ Var[\rho_i] &= \rho_i/\alpha_i \end{aligned}$$

For the case of multiple images, the maximum likelihood estimate is given by

$$\rho_i = \frac{X_{\Sigma_i}}{N\alpha_i}, \quad i = 1, \ldots, T \tag{7.9}$$

which has the following properties:

$$\begin{aligned} E[\rho_i] &= \rho_i \\ Var[\rho_i] &= \rho_i/(N\alpha_i) \end{aligned}$$

7.3.4 Confidence Intervals

Let $\chi = \{\mathbf{Z}_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_N\}$ denote a dataset of N images, with $\mathbf{Z}_k = (\mathbf{X}_k, \mathbf{Y}_k)$ containing the data for both border and non-border measurements for image k. Let $\mathbf{Z}_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_N$ be independent and identically distributed (i.i.d.) with distribution function F. Let F be the empirical distribution function of the observed data. Let $R(\chi, F)$ be a random vector giving the PSD estimated using either Miles-Lantuejoul or maximum likelihood. To construct confidence intervals for the estimated PSD, we require the distribution of $R(\chi, F)$. This distribution, called the sampling distribution, is unknown because F is unknown, being a function of the unknown PSD ρ. As $N \to \infty$, the limiting distribution of the maximum likelihood estimates is a multivariate normal with mean ρ and covariance $\mathbf{I}(\rho)^{-1}$, where $\mathbf{I}(\rho)$ is the Fisher information matrix, defined as

$$\mathbf{I}(\rho) = -E[\mathbf{l}''(\rho)]$$

in which $\mathbf{l}''(\rho)$ is a $T \times T$ matrix with the (i, j)th element given by $\frac{\partial^2 l(\rho)}{\partial \rho_i \partial \rho_j}$. Approximate confidence intervals for individual parameter estimates can be calculated as $\rho_i = \rho_i \pm \sigma_i z_\alpha$ in which σ_i is the ith diagonal element of the observed Fisher information matrix $-\mathbf{l}''(\rho)$ and z_α gives the appropriate quantile from the standard normal distribution for the confidence level α.

Given that the underlying distributions for $\mathbf{X}_k$ and $\mathbf{Y}_k$ are Poisson, we expect the sampling distributions to be non-normal in general. We therefore use bootstrapping [40, p.253] to approximate the distribution of $R(\chi, F)$ and construct confidence intervals. Let $\chi^* = \{\mathbf{Z}_1^*, \ldots, \mathbf{Z}_N^*\}$ denote a bootstrap sample of the dataset χ. The elements of χ^* are i.i.d. with distribution function F. In other words, χ^* is obtained by sampling χ N times where, for each of the N samples, the probability of selecting the data $\mathbf{Z}_k$ is 1/N. We denote a set of B bootstrap samples as $\chi_l^* = \{\mathbf{Z}_{l1}^*, \ldots, \mathbf{Z}_{lN}^*\}$ for $l = 1, \ldots, B$. The empirical distribution function of $R(\chi_l^*, F)$ for $l = 1, \ldots, B$ approximates the distribution function of $R(\chi, F)$, enabling confidence intervals to be constructed.

For $R(\chi, F) = \rho$, the distribution function of $R(\chi^*, F)$ is derived analytically using Equation (7.8). For $R(\chi, F) = \rho_b$ and $R(\chi, F) = \rho_{ML}$, the distribution of $R(\chi^*, F)$ is estimated using $B = 1000$ bootstrap samples. The confidence intervals are obtained using the percentile method, which consists of reading the appropriate quantiles from the cumulative distribution of $R(\chi^*, F)$. To calculate confidence intervals using the normal approximation, the observed Fisher information matrix $-\mathbf{l}''(\rho)$, also called the Hessian, must be calculated. The (i, j)th element of this matrix is given by

$$-\frac{\partial l(\rho)}{\partial \rho_i \partial \rho_j} = \frac{X_{\Sigma_i}}{\rho_i^2}\delta_{ij} + \sum_k^T \beta_{jk}\beta_{ik}\frac{Y_{\Sigma_k}}{m_{Y_k}^2} \tag{7.10}$$

7.4 Simulation methods

To investigate the performance of the maximum likelihood estimator relative to the standard Miles-Lantuejoul approach, these estimators were applied in several case studies. In each case study, 1000 simulations were carried out. Each simulation consists of generating a set of artificial images and applying the estimators to the particle length measurements obtained from these images. To generate an artificial image, we simulate the particle population in a local region surrounding the imaging volume V_I. In this region, we model the particle population as a three-dimensional stochastic process $\Phi = (\mathbf{X}_{wi}, L_i, \Theta_{zi})$ on $\mathbb{R}^3 \times \mathbb{R}^+ \times (-\pi/2, \pi/2]$ for $i = 1, \ldots, N_c$. $\mathbf{X}_{wi} = (X_{wi}, Y_{wi}, Z_{wi})$ gives the location of the centroid for particle i in the world coordinate frame, L_i gives the length, Θ_{zi} gives the orientation around the z-axis of the world coordinate frame, and N_c gives the number of particles. $\mathbf{X}_{wi}$, L_i, Θ_{zi}, and N_c are distributed independently of each other. X_{wi}, Y_{wi}, Z_{wi}, and Θ_{zi} are distributed uniformly on $[x_{\text{min}}, x_{\text{max}}]$, $[y_{\text{min}}, y_{\text{max}}]$, $[z_{\text{min}}, z_{\text{max}}]$, and $(-\pi/2, \pi/2]$, respectively. L_i has probability density function $p(L)$. N_c has a Poisson distribution with parameter $\lambda = N_c(x_{\text{max}} - x_{\text{min}})(y_{\text{max}} - y_{\text{min}})/ab$, in which N_c is the expected number of crystals per image, calculated from the PSD using

$$N_c = V_I \int_0^\infty f(l)dl$$

The size of the local region surrounding the imaging volume is defined by $(x_{\text{min}}, x_{\text{max}}) = (-0.5L_{\text{max}}, a + 0.5L_{\text{max}})$ and $(y_{\text{min}}, y_{\text{max}}) = (-0.5L_{\text{max}}, b + 0.5L_{\text{max}})$, in which L_{max} is defined as the size of the largest particle in the population. If L_{max} does not have a well-defined value, such as when the simulated density function $p(L)$ is normal, L_{max} is assigned the value corresponding to the 0.997th quantile of the simulated distribution function.

Each particle has a cylindrical geometry with height L_i and radius r, where r is assumed to be constant. Each particle is a convex, three-dimensional domain $P_i \in V$. To model the imaging process, P_i is projected onto an imaging plane using a camera model. This projection is computed by first applying rigid-body rotations and translations to change each point $\mathbf{X}_w$ in P_i from the world coordinate frame to the camera coordinate frame:

$$\mathbf{X_c} = \mathbf{R_z}\mathbf{R_y}\mathbf{R_x}\mathbf{X}_w + \mathbf{T} \tag{7.11}$$

in which $\mathbf{R_z}$, $\mathbf{R_y}$, and $\mathbf{R_x}$ are rigid-body rotation matrices, which are functions of the in-plane orientation θ_z and the orientations in depth θ_y and θ_x, respectively. $\mathbf{T} = (t_x, t_y, t_z)$ is a translation vector. Next, each point is projected onto the image plane according to some imaging model. Under perspective projection, the transformation from a 3-D point $\mathbf{X_c} = (X_c, Y_c, Z_c)$ in camera coordinates to an image point $\mathbf{x}_c = (x_c, y_c)$ is given by

$$x_c = \frac{f_c}{Z_c} X_c, \quad y_c = \frac{f_c}{Z_c} Y_c \tag{7.12}$$

in which f_c is the focal length of the camera. Figure S2 depicts the perspective projection of a cylindrical particle onto the image plane. Finally, to model CCD imaging, the image plane coordinates

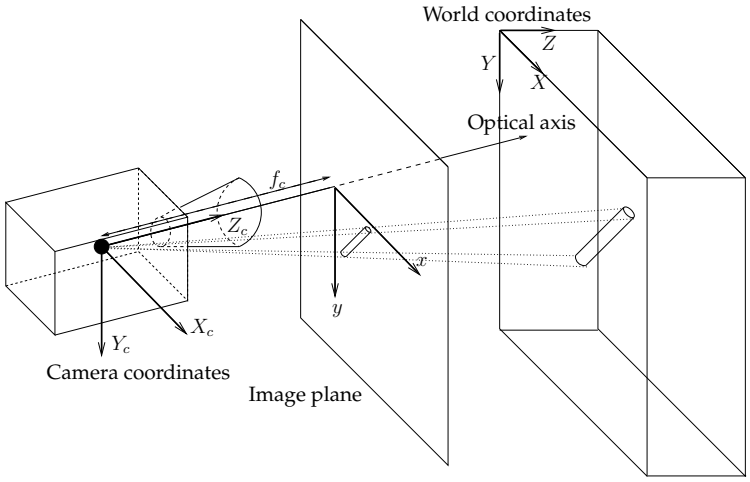


Figure S2: Depiction of the perspective projection of a cylindrical particle onto the image plane. For simplicity, the image plane is displayed in front of the camera.

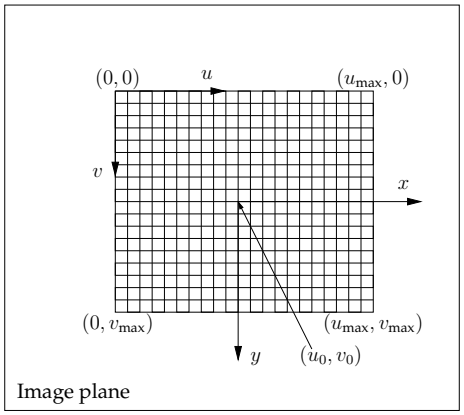


Figure S3: Depiction of CCD image.

$\mathbf{x}_c$ must be converted to pixel coordinates $\mathbf{w} = (u, v)$ using

$$u = u_0 + k_u x_c, \quad v = v_0 + k_v y_c \tag{7.13}$$

in which (u_0, v_0) corresponds to $\mathbf{x}_c = (0, 0)$ and k_u and k_v provide the necessary scaling based on pixel size and geometry. The CCD image is depicted in Figure S3. For our purposes, the projection of P_i onto the CCD array is simplified considerably by assuming the world coordinate frame and camera coordinate frame differ only by a translation in the z-direction. Thus, $X_c = X_w$ and $Y_c = Y_w$. Furthermore, the "weak perspective" projection model can be used because the depth of the imaging volume is small relative to the distance of the imaging volume from the camera. Therefore, f_c/Z_c and t_z can be assumed constant for all objects. Finally, we can assume that $(u_0, v_0) = (0, 0)$ and that the pixels are square such that $k_u = k_v$. Given these assumptions, the projection of a point $\mathbf{X}_w$ onto the CCD array is given simply by $(u, v) = (mX_w, mY_w)$, where

$m = k_u f_c / Z_c$. The length of an observed particle therefore equals its length in pixels divided by the factor m.

7.5 Results

Four different case studies were carried out to evaluate the effectiveness of the estimation methods. Figure S5 shows example images generated for each of these case studies. Each of these images has a horizontal image dimension of a=480 pixels and a vertical dimension of b=480 pixels. The first row displays four simulated images for monodisperse particles of length $0.5a$ with N_c=25 crystals per image. The second row shows images of particles uniformly distributed on [$0.1a$ $0.9a$] with N_c=25. The third row shows images of particles normally-distributed with $\mu = 0.5a$ and $\sigma = 0.4a/3$ with N_c=25, and the fourth row shows example images for simulations of particles uniformly-distributed on [$0.1a$ $2.0a$] with N_c=15.

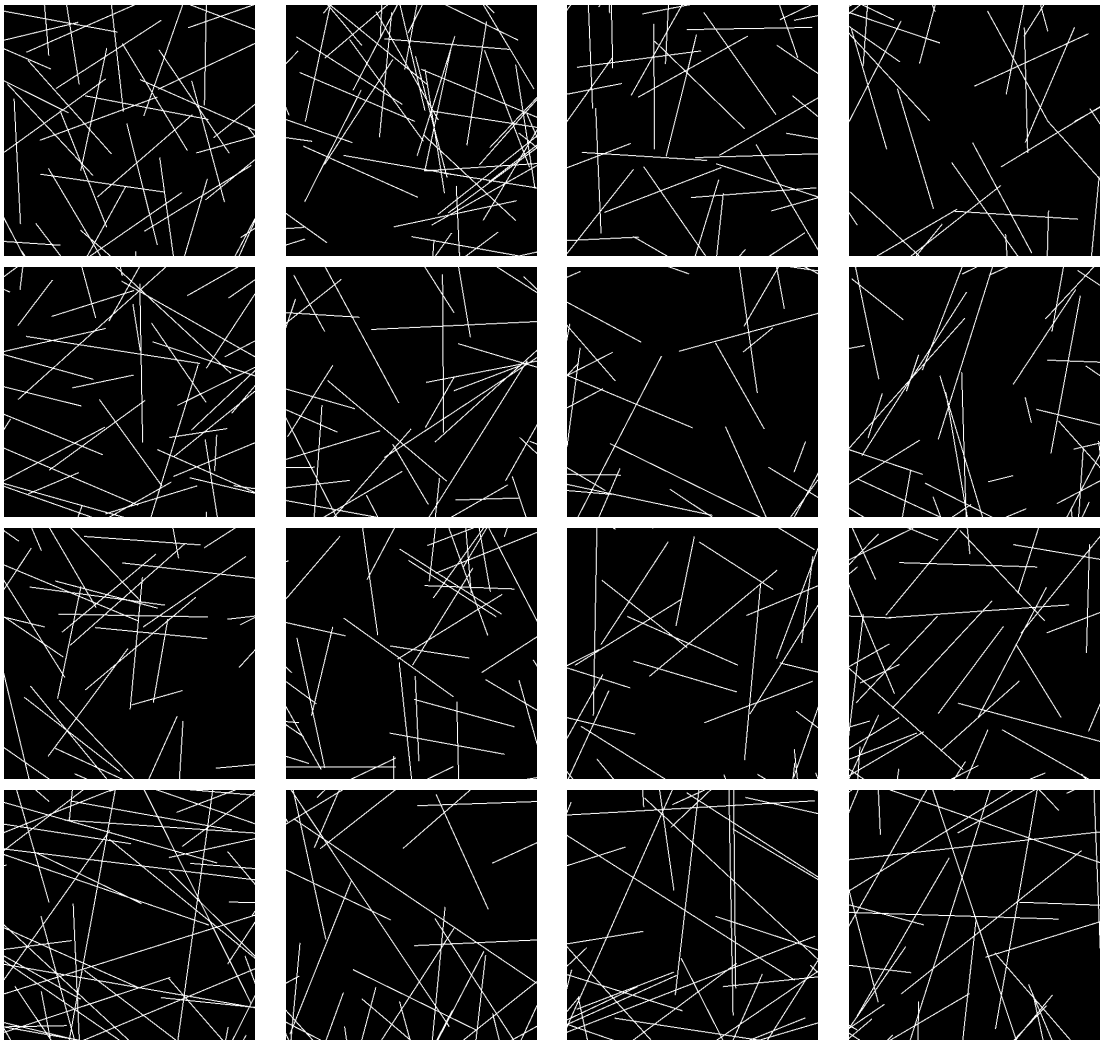
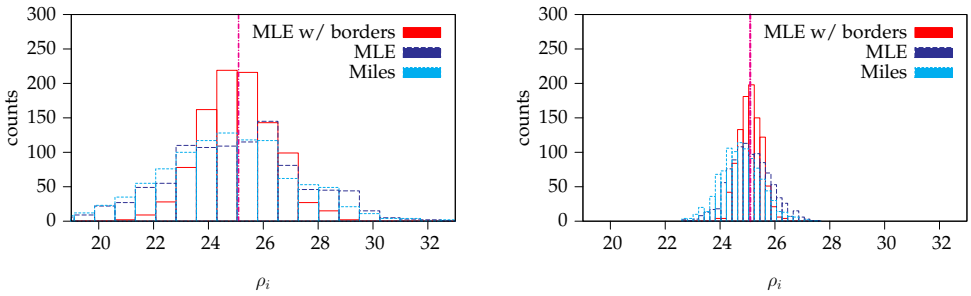


Figure S4: Example images for simulations of various particle populations. Row 1: monodisperse particles of length $0.5a$, N_c=25. Row 2: particles uniformly distributed on [$0.1a$ $0.9a$]. Row 3: particles normally-distributed with $\mu = 0.5a$ and $\sigma = 0.4a/3$, N_c=25. Row 4: particles uniformly-distributed on [$0.1a$ $2.0a$], N_c=15.



(a) (b)

Figure S5: Comparison of estimated sampling distributions for absolute PSD for monodisperse particles. Results based on 1000 simulations, 10 size classes, N_c=25. (a) Results for 10 images/simulation. (b) Results for 100 images/simulation.

7.5.1 Case study 1: monodisperse particles of length 0.5a

In the first case study, the particle population consists of monodisperse particles of length 0.5a. The first row in Figure S5 shows example images from these simulations. The length scale is discretized on [0.1a $\sqrt{2}a$] into T=10 size classes with the fourth size class centered at 0.5a. The sampling distributions for the various estimators are shown in Figure S5 for 1000 simulations using 10 images/simulation and 100 images/simulation. Including the border particle measurements provides better estimates, as evidenced by the lower variance in the sampling distribution for ρ_b relative to the other estimators. As a measure of the improvement gained by including the border particle measurements, we calculate the relative efficiency of ρ_{b_i} versus ρ_{ML_i} for a given size class i as

$$\text{eff}(\rho_{b_i}, \rho_{ML_i}) = \frac{\text{MSE}(\rho_{ML_i})}{\text{MSE}(\rho_{b_i})} \tag{7.14}$$

in which $\text{MSE}(T) = var(T) + [\text{bias}(T)]^2 = E[(T-\rho)^2]$ is the mean-squared error for estimator T. The MSE is estimated for size class i as

$$\text{MSE}(T_i) = \frac{1}{n}\sum_{j=1}^{n}(T_j - \rho_i)^2 \tag{7.15}$$

in which n is the number of simulations. The relative efficiency of the estimators appears relatively independent of the number of images per simulation, with values ranging between 3.5 and 4.0 as the number of images per simulation is varied between 10 and 100. Thus, for this particular case, including the border particle measurements decreases the number of images required to obtain a given accuracy by a factor of about four. For monodisperse systems in general, we would expect the efficiency to be a monotonically increasing function of particle size.

Figure S6 demonstrates the effectiveness of the bootstrap approach for determining confidence intervals. Figure S6 is constructed by calculating bootstrap confidence intervals for 1000 different simulations and determining what fraction of these confidence intervals contain the true parameters for a given level of confidence, or α. The figure shows results for confidence intervals

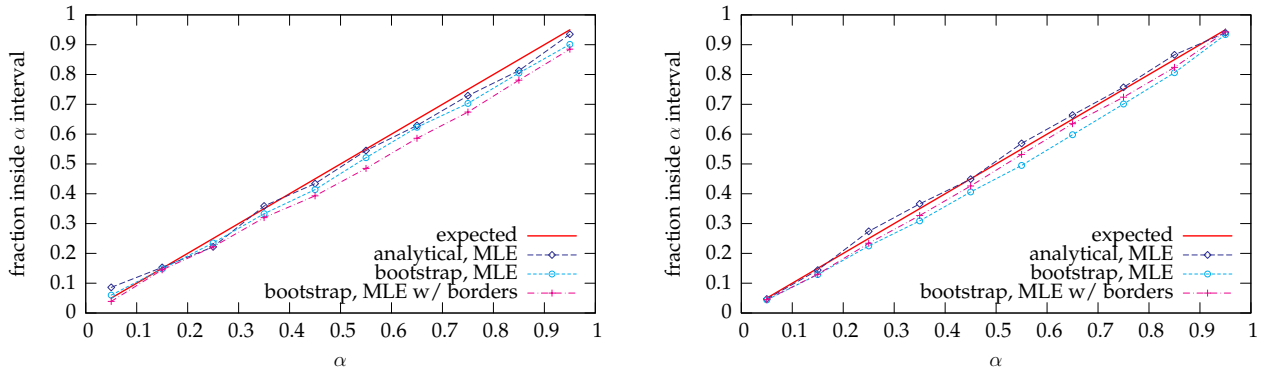


Figure S6: Fraction of confidence intervals containing the true parameter value versus confidence level. Results based on 1000 simulations, 10 size classes (results shown only for size class corresponding to monodisperse particle size), $N_c = 25$. (a) Results for 10 images/simulation. (b) Results for 100 images/simulation.

based on the analytical sampling distribution of ρ (Equation (7.8) as well as sampling distributions estimated using bootstrapping for both ρ_b and ρ. The fraction of confidence intervals containing the true parameter corresponds closely to the expected value (i.e. the confidence level), even for the case of only 10 images/simulation.

7.5.2 Case study 2: uniform distribution on [0.1a 0.9a]

In the second case study, the particle population consists of particles uniformly distributed on [0.1a 0.9a]. The second row in Figure S5 shows example images from these simulations. The length scale is discretized on [0.1a 0.9a] into T=10 size classes of equal size.

The efficiency of ρ_b relative to ρ_{ML}, calculated using Equation (7.14), is plotted versus size class in Figure S7. This plot indicates that including the border particle measurements does not appear to improve the estimation for the lower size classes but results in a significant increase in efficiency for the largest size class.

Figures S8 and S9 plot the fraction of bootstrap confidence intervals containing the true value of ρ_i for various size classes based on 100 images/simulation and 10 images/simulation. The bootstrap approach is effective for 100 images/simulation but underestimates the size of the confidence interval for 10 images/simulation. Including the border particle measurements enables better determination of confidence intervals, particularly for the largest size class.

7.5.3 Case study 3: normal distribution

For the third case study, the particle population consists of particles with lengths distributed as a normal with $\mu = 0.5a$ and $\sigma = 0.4a/3$. The third row in Figure S5 shows example images from these simulations. The length scale is discretized on $[\mu - 3\sigma, \mu + 3\sigma]$=[0.1$a$ 0.9a] into T=10 equi-

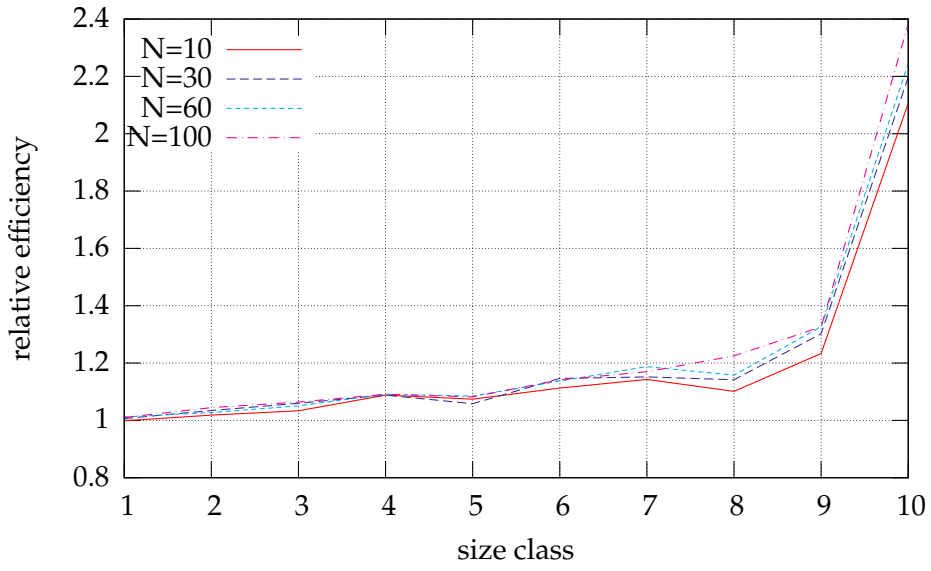


Figure S7: Relative efficiencies (eff(ρ_{b_i}, ρ_{ML_i})) plotted versus size class for various numbers of images per simulation: case study 2.

spaced size classes. Figure S10 illustrates the sampling distributions at the various size classes for ρ_i. The x-y plane of Figure S10 shows the histogram generated for a normal distribution. The discrete sampling distributions, calculated using Equation (7.8), are plotted for each size class. This figure indicates that the sampling distribution for a given size class can be adequately represented by a normal distribution provided the density of particles in that size class is sufficiently high. However, for the larger size classes, approximating the sampling distribution as a normal would lead to inaccurate confidence intervals.

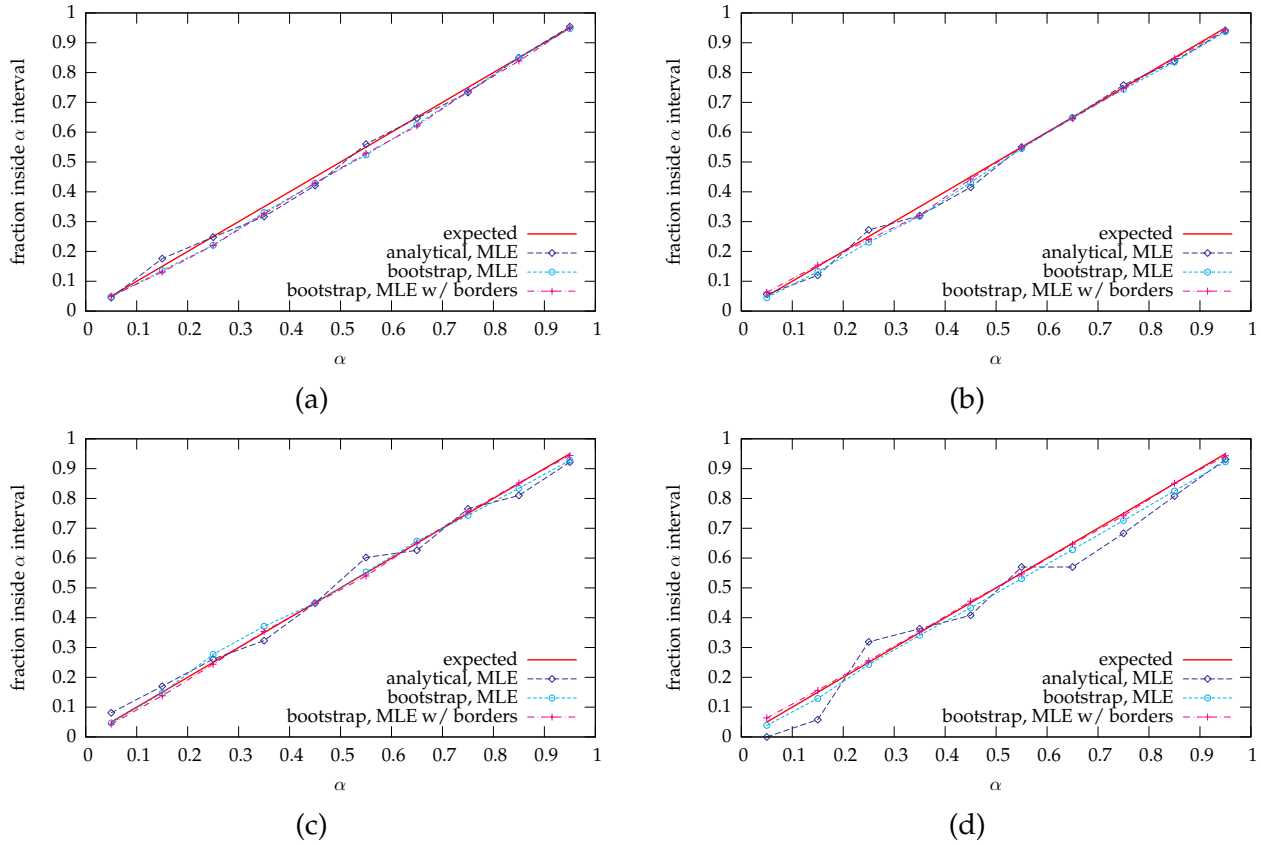


Figure S8: Fraction of confidence intervals containing true parameter values for different confidence levels, N=100. (a) Size class 1 (smallest size class). (b) Size class 4. (c) Size class 7. (d) Size class 10 (largest size class).

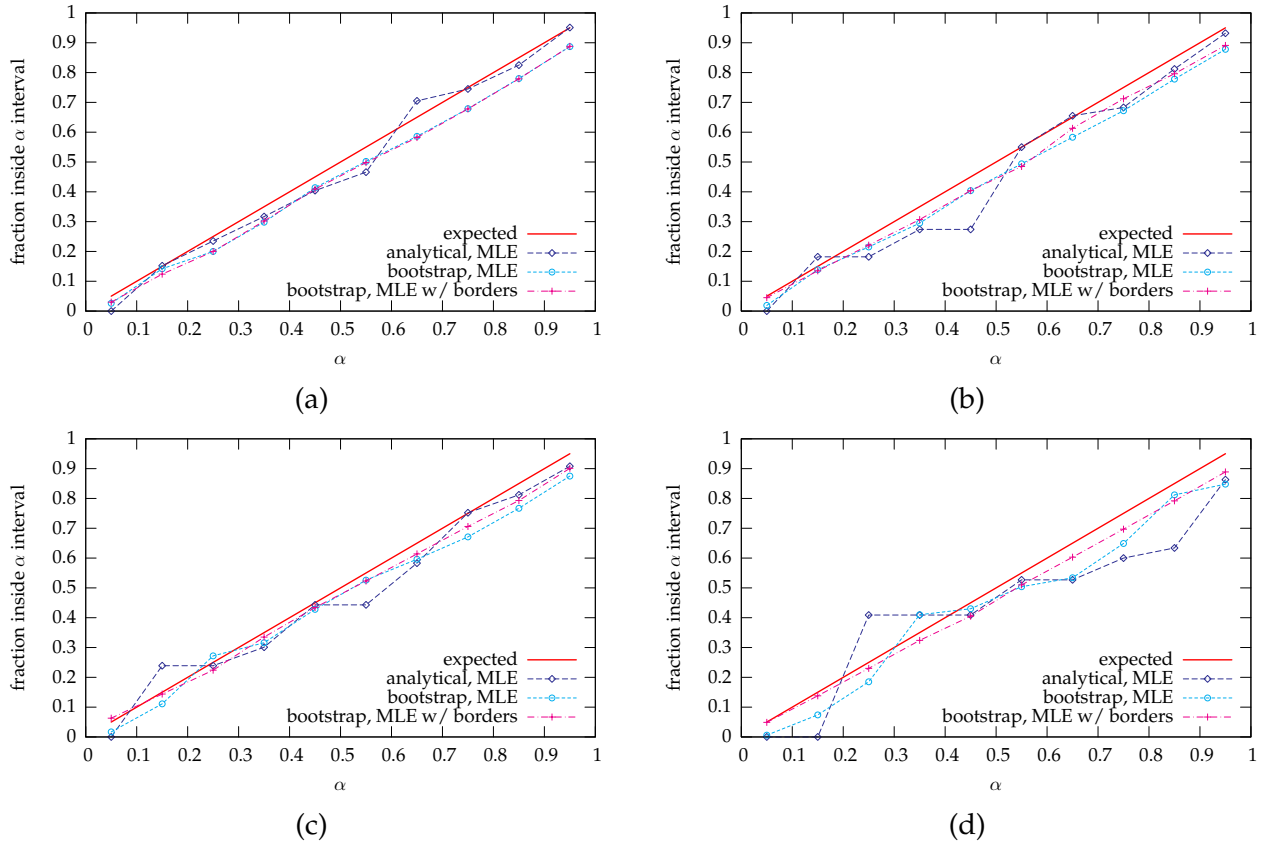


Figure S9: Fraction of confidence intervals containing true parameter values for different confidence levels, N=10. (a) Size class 1 (smallest size class). (b) Size class 4. (c) Size class 7. (d) Size class 10 (largest size class).

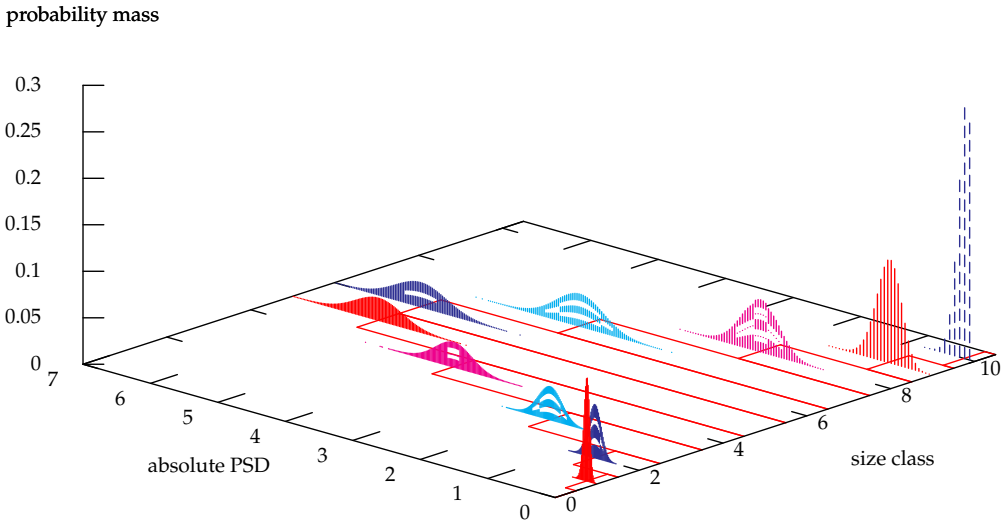


Figure S10: Sampling distributions for the various size classes of a discrete normal distribution. $N = 100$.

Figure S11 plots eff$(\rho_{b_i}, \rho_{ML_i})$ for various numbers of images per simulation. Comparing Figure S11 with Figure S7 indicates that the relative efficiency for a given size class is a function of both the size and the density of the particles in that size class.

7.5.4 Case study 4: uniform distribution on [0.4a 2.0a]

In the fourth case study, the particle population consists of particles uniformly distributed on [0.4a 2.0a]. The fourth row in Figure S5 shows example images from these simulations. The length scale was discretized on [0.4a a] into $T - 1 = 9$ bins with the Tth bin extending from a to $L_{\max}$. $L_{\max}$ was assumed unknown and estimated with initial value $\sqrt{2}a$. That is, Equation (A.3) was solved as before with the exception that the parameters m_{X_i} and m_{Y_j} were updated at each iteration based on the current estimate of $L_{\max}$. Figure S12 shows the sampling distributions for ρ_i for various size classes, as well as the sampling distribution for $L_{\max}$. The MLE w/ borders approach is effective in estimating $L_{\max}$ and the density of particles in the largest size class. It should be remembered, however, that the estimation is based on the assumption that the particles are uniformly distributed across each size class. This assumption is suitable for finely discretized size classes but is probably not suitable for the single, large, oversized size class. Thus, the estimated value of $L_{\max}$ should be used only as a rough estimate of the true value and as an indication for

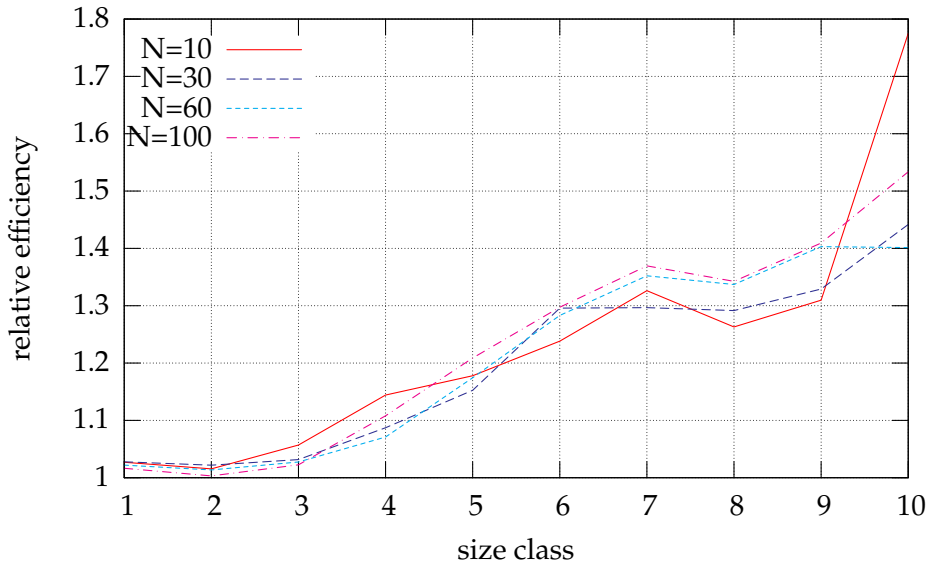


Figure S11: Relative efficiencies (eff(ρ_{b_i}, ρ_{ML_i})) plotted versus size class for various numbers of images per simulation: case study 3.

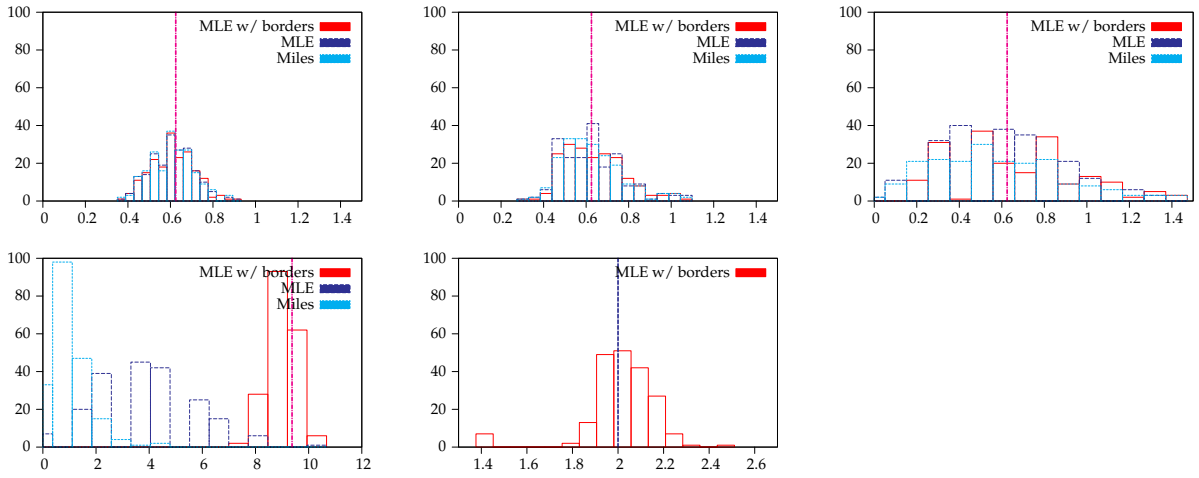


Figure S12: Comparison of sampling distributions for absolute PSD for particles distributed uniformly on [0.4a 2.0a]. Results based on 200 simulations, 100 images/simulation, 10 size classes, N_c=15. (a) Size class 1. (b) Size class 5. (c) Size class 9. (d) Size class 10. (e) $L_{\max}$.

the appropriateness of the camera magnification.

7.6 Conclusion

The maximum likelihood estimator for imaging-based PSD measurement of zero-width, needle-like particles has been derived using both censored and uncensored observations (i.e. border and non-border particles). The performance of the estimator has been compared with the standard Miles-Lantuejoul approach using four case studies that highlight several advantages of the MLE approach. The case studies indicate that MLE is more efficient than Miles-Lantuejoul, particularly if the particle population is monodisperse or contains particles that are large relative to the size of the image. Furthermore, MLE can estimate the number density of over-sized particles (particles bigger than the image dimension) along with the size $L_{\max}$ of the largest particle while the Miles-Lantuejoul approach can be applied only for particles smaller than the image dimension.

The limitations of the MLE approach should also be discussed. The primary limitation of the MLE derived in this paper is due to the assumption that the particles have needle-like geometry. The Miles-Lantuejoul approach, on the other hand, can be applied to a much wider class of geometries. Secondly, the MLE approach requires the solution of a nonlinear optimization problem. Thus, confidence interval determination by bootstrapping can be computationally-intensive. Finally, it should be noted that the MLE estimates related to over-sized particles are obtained by making the rather unrealistic assumption that over-sized particles are uniformly distributed in length on $[a\ L_{\max}]$. The estimates related to over-sized particles are therefore biased in general but may be useful for identifying whether or not the camera magnification is suitable for the given system.

Several areas for future work are evident. Choosing the optimal number, location, and size of bins for constructing histograms should be addressed. Integrating measurements taken at multiple scales or magnifications is also important. For systems of high-aspect-ratio particles, incorporating the width of border particles into the estimation could lead to increased efficiency by narrowing down the number of size classes to which a border particle may correspond. Methods for estimating the PSD when occlusion or overlap effects are not negligible are necessary for systems at high solids concentrations.

7.7 Acknowledgment

This material is based upon work supported by the National Science Foundation under Grant No. 0540147.

Chapter 8

Assessing the reliability of imaging-based, number density measurement

8.1 Introduction

Advanced control of crystal shape, size distribution, and polymorphic form (internal lattice structure) in suspension crystallization processes has been hindered by the limitations of available online sensors. High-speed, in situ video microscopy is a promising technology for measuring these critical solid-phase properties. Several challenges must be addressed to use in situ microscopy for PSD measurement. The first challenge is achieving image *segmentation*, or separating the objects of interest (e.g. the particles) from the background. Methods for segmenting in situ images have been developed for circular particles [99], arbitrarily-shaped particles [19], elliptical particles [47], needle-like particles [59], and particles with shape that can be represented using a wireframe model [60].

The second challenge is to estimate the PSD given the size and shape information obtained through successful image segmentation. The estimation is complicated by the presence of two types of edge effects: sampling bias and censoring. Sampling bias occurs when the probability of observing an object depends on its size or shape. Sampling bias accompanies imaging-based PSD measurement because particles that are large relative to the size of the image have a high probability of intersecting the image boundary while smaller particles have a low probability of intersecting the image boundary. Censoring occurs when only partial information is obtained for a given observation. For example, an observation of a particle touching the image border reveals that the particle is at least as long as the observed length, but the true length is unknown. Miles [72] and Lantuejoul [57] derived an unbiased estimator of PSD that corrects sampling bias using only uncensored observations. We [**?**] derived the maximum likelihood estimator for the PSD of high-aspect-ratio particles including censored observations. Both of these methods assume the objects of interest are resting in a plane perpendicular to the optical axis of the camera. If the particles are randomly oriented in three-dimensional space, the two-dimensional projected lengths are less than or equal to the true lengths, resulting in another type of censoring for which these previous methods do not account. Baddeley [5] provides an excellent review of various methods for correcting edge effects under a variety of situations.

A third challenge lies at the interface of the two challenges just mentioned. The methods for PSD estimation in the presence of sampling bias and censoring are based on the assumption that image segmentation is perfect, or that every single particle appearing in the image is identified correctly. Achieving perfect image segmentation is a realistic goal, however, only at low solids concentrations. At higher solids concentrations, a significant fraction of the particles may be overlapping or occluded, which often results in these particles being missed by the automated image analysis algorithm. The density of particles is thus underestimated. The objective of this work is to address this third challenge by developing a semi-empirical, probabilistic model that enables PSD estimation for imperfect image analysis. This paper also presents a descriptor that can be used to quantify the expected amount of particle overlap and thereby assess the reliability of the PSD estimate.

The paper is organized as follows. Section 8.2 presents previous work relevant to the scope of this paper. Section 8.3 presents theory related to particle overlap probabilities, proposes a descriptor for measurement reliability, and presents an estimator for particle number density that accounts for particle overlap. Section 8.4 describes the simulation studies used to determine the conditions for which SHARC gives an accurate PSD measurement. Section 8.5 presents the results of these simulation studies, and Section 8.6 summarizes our findings.

8.2 Previous Work

Armitage [4] was one of the first to investigate the effects of particle overlap on particle number density estimation. Armitage derived formulas for the expected numbers of isolated clumps (groups of one or more overlapping particles) for circular and rectangular particles, and used these formulas to estimate the mean clump size and number density of particles. Mack [68, 69] extended Armitage's results to three-dimensional particles of any convex shape, deriving formulas for the expected numbers of clumps and isolated particles based on the perimeters and areas of the two-dimensional projections of the particles. Roach [91] summarized the work of Mack and Armitage, described several applications, and developed theory for specialized applications. Kellerer [56] was the first to account for edge effects and overlap simultaneously and derived a formula for the expected number of clumps minus the number of enclosed voids (i.e. regions of the image completely enclosed by particles). To model the complicated, random shapes that form due to the random placement and overlap of simple geometric objects, the Boolean model [1, 6] can be used.

The methods cited above assume that a clear distinction can be made between the image background and the objects of interest, or that the image can be segmented by simple thresholding. In some applications, however, more advanced image analysis is required, and modeling the output of a complicated image analysis algorithm to enable statistical inference is non-trivial. For example, we previously developed a novel image analysis algorithm called SHARC that automatically extracts particle size information from in situ images of needle-like crystals [59]. Using images acquired during a pharmaceutical crystallization experiment, we demonstrated that SHARC's PSD measurements are consistent with measurements obtained through manual, human analysis of the images, provided the images are acquired under suitable conditions. As a

model-based vision algorithm, SHARC can successfully identify overlapped particles provided the degree of overlap is minor. Initial studies have shown that the methods cited above do not adequately model the output of the SHARC algorithm. In this work, we extend these methods to enable statistical inference for SHARC's output. We expect the methodology developed here to be applicable to other situations in which advanced image analysis is necessary to extract information from noisy images.

8.3 Theory

8.3.1 Particulate system definition

Consider a slurry of volume V in which a solid phase of discrete particles is dispersed in a continuous fluid phase. Let ρ_c be the solid phase density (mass solid/volume solid) and M_T the slurry density (mass solid/volume slurry), or solids concentration. The volume solids concentration (volume solid/volume slurry) is given by M_T/ρ_c. Let L be the characteristic length of a particle and define a shape factor k_v such that the volume of a single particle is given by $k_v L^3$. Let $f(L)$ denote the PSD, or the number of particles of characteristic length L per unit volume slurry. The ith moment of the PSD is denoted $\mu_i = \int_0^\infty f(L)L^i dL$. Given the PSD, the slurry density can be calculated as $M_T = \rho_c k_v \mu_3$ assuming k_v is independent of length. The zeroth moment, μ_0, equals the number of particles per unit volume slurry. For a monodisperse system in which all particles have the same length l, $\mu_0 = M_T/\rho_c k_v l^3$.

8.3.2 Sampling and measurement definitions

Let $V_I \in V$ denote an imaging volume, and let I denote an image created by perspective projection of V_I onto a two-dimensional image plane. Let a and b denote the horizontal and vertical dimensions of V_I, or the field of view, and let d denote the depth dimension of V_I, or the depth of field. Thus, the volume of V_I is given by abd, and the average number of particles per image is given by $N_c = \mu_0 abd$.

Number-based PSDs are typically measured by discretizing the characteristic length scale into T non-overlapping bins or size classes. We therefore define the discrete PSD as

$$\rho_i = \int_{S_i}^{S_{i+1}} f(l)dl, \quad i = 1, \ldots, T \tag{8.1}$$

in which $\mathbf{S} = (S_1, \ldots, S_{T+1})$ is the vector of breaks between size classes. In this work, we consider only monodisperse populations, so we represent the PSD for a given system using a single scalar value ρ. For a monodisperse system, ρ equals the number density.

8.3.3 Descriptor for number density reliability

Our goal is to estimate the number density of particles, and our hypothesis is that the quality or reliability of the estimate will correlate with the amount of particle overlap. We therefore want

to quantify the amount of particle overlap observed in the acquired images. We can calculate the probability that a given particle will be overlapped by other particles as follows. Consider a population of n identical particles randomly located within a system volume V_S. The number density of particles is given by n/V_S. Let the particles be projected orthogonally onto a plane of area A_S, giving a particle density per unit area of $\rho_A = n/A_S$. Let K be a random variable giving the number of times a given particle's projection is overlapped by other particles' projections. Assuming A_S is sufficiently large that edge effects are negligible, the probability that the projection of a given particle will be overlapped by the projection of a second, given particle is $p_{\text{ovp}} = \Omega/A_S$, in which Ω denotes the admissible area, or the area of the region inside which the second particle's projection overlaps the first particle's projection. Thus, the probability that a given particle will be overlapped by k particles is given by the binomial distribution with n Bernoulli trials and the probability of success in each trial given by p_{ovp},

$$p_K(k) = \begin{pmatrix} n \\ k \end{pmatrix} p_{\text{ovp}}^k (1 - p_{\text{ovp}})^{n-k} \tag{8.2}$$

As $n \to \infty$ with $\rho_A = n/A_S$ constant, the binomial distribution converges to the Poisson distribution with parameter $\lambda = \rho_A \Omega$.

$$p_K(k) = \frac{e^{-\lambda} \lambda^k}{k!} \tag{8.3}$$

Thus, the probability that a given particle will be completely isolated is given by $p_{\text{iso}} = p_K(0) = \exp(-\rho_A \Omega)$.

As already mentioned, we expect the reliability of the PSD estimate to correlate with particle overlap. Therefore, as an indicator of the reliability of the PSD estimate, we define the parameter D as

$$D = -\log(p_{\text{iso}}) = \rho_A \Omega \tag{8.4}$$

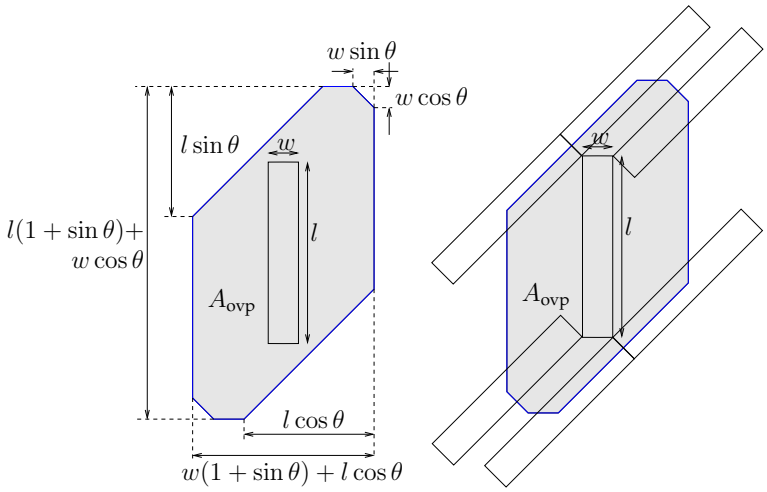
The area density ρ_A can be calculated as $\rho_A = \mu_0 d$. The admissible area Ω depends on the particle geometry. To illustrate how the admissible area can be calculated, consider the rectangular particle of length l and width w in the center of Figure S1. Next, consider a second particle of identical dimensions placed in the image with orientation θ. If the midpoint of the second particle is placed anywhere inside the shaded region in Figure S1, the first and second particles overlap. This fact is illustrated in Figure S1(b) using the various particles placed around the border of the shaded area. Using straightforward geometric relationships (see Figure S1(a)), the area A_{ovp} of this shaded region can be shown to be

$$A_{\text{ovp}} = (w(1 + \sin\theta) + l\cos\theta)(l(1 + \sin\theta) + w\cos\theta) - l^2 \cos\theta \sin\theta - w^2 \cos\theta \sin\theta \tag{8.5}$$

and Ω is obtained by integrating over θ

$$\Omega = E[A_{\text{ovp}}] = \frac{1}{\pi} \int_{-\pi/2}^{\pi/2} A_{\text{ovp}} d\theta = \frac{2}{\pi} \left(l^2 + w^2 + lw(2 + \pi) \right) \tag{8.6}$$

in which a uniform distribution in orientation has been assumed.



(a) (b)

Figure S1: Geometric representation of region in which particles of specific orientation and shape overlap.

Mack [68] (see also Roach [91, p.44]) derived a more general result that enables the calculation of the admissible area Ω for any set of convex bodies of identical size and shape and having random orientation. Mack's surprisingly simple result gives Ω based on the area a_p and perimeter s_p of a two-dimensional domain:

$$\Omega = 2a_p + \frac{s_p^2}{2\pi} \tag{8.7}$$

8.3.4 Estimation of number density

Given our hypothesis that particle overlap is the primary cause of failure for image analysis-based measurement, it seems reasonable to estimate the number density of particles based on the number of completely isolated particles observed in the image. Letting X be a random variable giving the number of observations of completely isolated, non-border particles, it can be shown that the probability density for X is Poisson with parameter $m_X = \rho_A N A_{NB} \exp(-\rho_A \Omega)$, in which N is the number of images, A_{NB} is the area of the region inside which a particle does not touch the image border, Ω is the admissible area defined above, and ρ_A is the area number density as defined above. The maximum likelihood estimate of ρ_A is therefore

$$\rho_A = \arg\max_{\rho_A} \frac{e^{-m_X} m_X^x}{x!} \tag{8.8}$$

in which x is the realization of X, or the measured number of isolated, non-border particles.

The estimator given by Equation (8.8) is correct if the image analysis algorithm identifies only isolated particles. Model-based image analysis algorithms, however, are designed to identify the objects of interest even in the presence of overlap or occlusion. The number of particles observed by a model-based image analysis algorithm is therefore greater than the number

of isolated particles, and Equation (8.8) can be expected to give bad estimates given such data. Assuming the number of particles identified by any given image analysis algorithm depends primarily on the amount of overlap, a reasonable model can be formulated based on the overlap model given above. Letting X denote the number of particles identified by an image analysis algorithm, a reasonable form for the probability density of X would be Poisson with parameter $m_X = \rho_A N A_{NB} \exp(-\rho_A \Omega \theta)$, in which θ is an empirical parameter and the other variables are as defined previously.

8.4 Simulation methods

This section describes the methods used to generate artificial images and analyze the images.

8.4.1 Artificial image generation

To generate an artificial image, we simulate the particle population in a local region surrounding the imaging volume V_I. In this region, we model the particle population as a three-dimensional stochastic process $\Phi = (\mathbf{X}_{wi}, L_i, \Theta_{zi})$ on $\mathbb{R}^3 \times \mathbb{R}^+ \times (-\pi/2, \pi/2]$ for $i = 1, \ldots, N_c$. $\mathbf{X}_{wi} = (X_{wi}, Y_{wi}, Z_{wi})$ gives the location of the centroid for particle i in the world coordinate frame, L_i gives the length, Θ_{zi} gives the orientation around the z-axis of the world coordinate frame, and N_c gives the number of particles. $\mathbf{X}_{wi}$, L_i, Θ_{zi}, and N_c are distributed independently of each other. X_{wi}, Y_{wi}, Z_{wi}, and Θ_{zi} are distributed uniformly on $[x_{\min}, x_{\max}]$, $[y_{\min}, y_{\max}]$, $[z_{\min}, z_{\max}]$, and $(-\pi/2, \pi/2]$, respectively. L_i has probability density function $p(L)$. N_c has a Poisson distribution with parameter $\lambda = N_c(x_{\max} - x_{\min})(y_{\max} - y_{\min})/ab$, in which N_c is the expected number of crystals per image, calculated from the PSD using

$$N_c = V_I \int_0^\infty f(l) dl$$

The size of the local region surrounding the imaging volume is defined by $(x_{\min}, x_{\max}) = (-0.5L_{\max}, a + 0.5L_{\max})$ and $(y_{\min}, y_{\max}) = (-0.5L_{\max}, b + 0.5L_{\max})$, in which $L_{\max}$ is defined as the size of the largest particle in the population. If $L_{\max}$ does not have a well-defined value, such as when the simulated density function $p(L)$ is normal, $L_{\max}$ is assigned the value corresponding to the 0.997th quantile of the simulated distribution function.

Each particle is a convex, three-dimensional domain $P_i \in V$. To model the imaging process, P_i is projected onto an imaging plane using a camera model. This projection is computed by first applying rigid-body rotations and translations to change each point $\mathbf{X}_w$ in P_i from the world coordinate frame to the camera coordinate frame:

$$\mathbf{X_c} = \mathbf{R_z R_y R_x X}_w + \mathbf{T} \tag{8.9}$$

in which $\mathbf{R_z}$, $\mathbf{R_y}$, and $\mathbf{R_x}$ are rigid-body rotation matrices, which are functions of the in-plane orientation θ_z and the orientations in depth θ_y and θ_x, respectively. $\mathbf{T} = (t_x, t_y, t_z)$ is a translation vector. Next, each point is projected onto the image plane according to some imaging model.

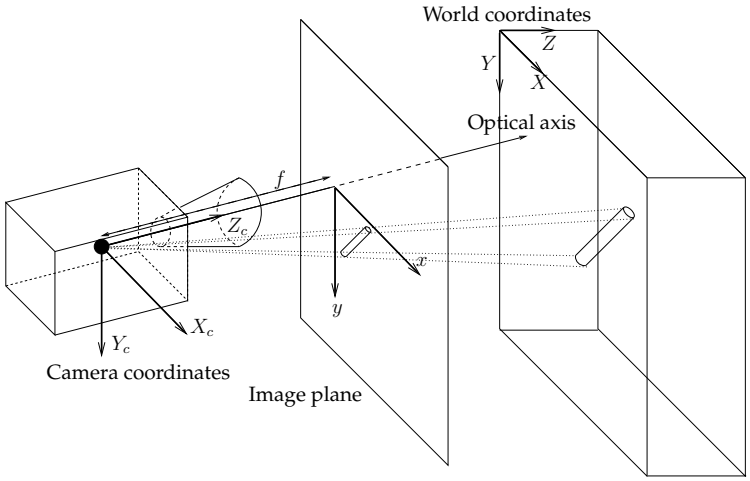


Figure S2: Depiction of the perspective projection of a cylindrical particle onto the image plane. For simplicity, the image plane is displayed in front of the camera.

Under perspective projection, the transformation from a 3-D point $\mathbf{X_c} = (X_c, Y_c, Z_c)$ in camera coordinates to an image point $\mathbf{x}_c = (x_c, y_c)$ is given by

$$x_c = \frac{f_c}{Z_c} X_c, \quad y_c = \frac{f_c}{Z_c} Y_c \tag{8.10}$$

in which f_c is the focal length of the camera. Figure S2 depicts the perspective projection of a cylindrical particle onto the image plane. Finally, to model CCD imaging, the image plane coordinates $\mathbf{x}_c$ must be converted to pixel coordinates $\mathbf{w} = (u, v)$ using

$$u = u_0 + k_u x_c, \quad v = v_0 + k_v y_c \tag{8.11}$$

in which (u_0, v_0) corresponds to $\mathbf{x}_c = (0, 0)$ and k_u and k_v provide the necessary scaling based on pixel size and geometry. The CCD image is depicted in Figure S3. For our purposes, the projection of P_i onto the CCD array is simplified considerably by assuming the world coordinate frame and camera coordinate frame differ only by a translation in the z-direction. Thus, $X_c = X_w$ and $Y_c = Y_w$. Furthermore, the "weak perspective" projection model can be used because the depth of the imaging volume is small relative to the distance of the imaging volume from the camera. Thus, f_c/Z_c and t_z can be assumed constant for all objects. Finally, we can assume that $(u_0, v_0) = (0, 0)$ and that the pixels are square such that $k_u = k_v$. Given these assumptions, the projection of a point $\mathbf{X}_w$ onto the CCD array is given simply by $(u, v) = (mX_1, mY_1)$, where $m = k_u f_c / Z_c$.

The information required to generate an artificial image includes the image dimensions a and b and magnification m, the length distribution $p(L)$, the aspect ratio AR, and the number of particles per image λ. For simulations at a given solids concentration M_T, λ is calculated as

$$\lambda = \frac{M_T abd}{\rho_c \int_0^\infty p(L) k_v L^3 dL}$$

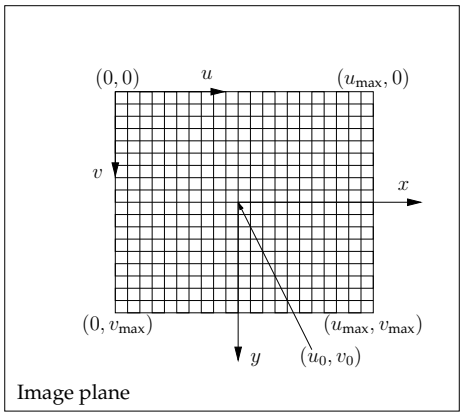


Figure S3: Depiction of CCD image.

Description	Symbol	Value
Horizontal dimension of imaging volume	a	2 mm
Vertical dimension of imaging volume	b	2 mm
Depth of field	d	0.25 mm
Solid phase density	ρ_c	2 mg/mm^3
Number of horizontal CCD pixels	$u_{\max}$	480
Number of vertical CCD pixels	$v_{\max}$	480

Table 8.1: Parameters used to simulate imaging of particle population at a given solids concentration.

For a monodisperse population with $p(L) = \delta(L - l)$, λ is given by

$$\lambda = \frac{M_T abd}{\rho_c k_v l^3}$$

For simulations at a given value of D, λ is calculated using Equation (8.4):

$$\lambda = D/p_{\text{ovp}}(L) + 1$$

in which L denotes the mean particle length.

8.4.2 Justifications for two-dimensional system model

The assumptions justifying our use of a two-dimensional process to model a three-dimensional system are as follows. First of all, we assume the camera is positioned a fixed distance z_0 from the imaging volume, and that $d \ll z_0$. This assumption means the particles in the imaging volume are projected onto the image plane according to the weak perspective projection model. In other words, the projected particle lengths measured in the image coordinate system can be related to the true projected particle lengths by applying a constant magnification factor m, without regard

Table 8.2: Parameter values used to analyze images from crystallization experiment.

Line finder Parameters		Collinearity Thresholds		Parallelism Thresholds	
n_∇	5	ϵ_{θ_C}	20 degrees	ϵ_{θ_P}	5 degrees
$\epsilon_{\lvert\nabla\rvert}$	1	ϵ_{EP}	0.5	ϵ_Q	0.85
n_b	6 buckets	ϵ_{PD}	0.5	ϵ_{AR}	4.5
ϵ_A	20 pixels				

for the distance of the particle from the camera. Secondly, we assume all particles are oriented in a plane orthogonal to the camera's optical axis. This assumption, together with the weak perspective assumption, essentially reduces the 3-D process to a 2-D process, thereby simplifying the analysis considerably. These assumptions are not used only for convenience, however, but rather to reflect the actual conditions under which in situ imaging measurements are made in practice. To obtain useful in situ images in high solids concentrations, the camera must have a small depth of field and be focused only a small depth into the particulate slurry. It seems reasonable, therefore, to expect the shear flow at the slurry-sensor interface to cause the particles to align orthogonal to the this interface, and thus orthogonal to the camera's optical axis.

8.4.3 Image analysis methods

To determine the conditions under which reliable measurements can be obtained using image analysis, we applied the SHARC algorithm described in [59] to artificial images generated at various D. SHARC is designed to identify high-aspect-ratio particles in the presence of particle overlap, and its effectiveness for real images relative to manual image analysis by human operators has been established previously [59].

SHARC was applied to the artificial images using the parameters shown in Table 8.2.

8.5 Results

8.5.1 Descriptor comparison: M_T versus D

Figure S4 compares images generated for two different monodisperse particle populations generated at the same solids concentration M_T and at the same D. The images generated at constant D appear to be more similar with respect to the amount of particle overlap than the images generated at constant solids concentration. This qualitative assessment is confirmed by Figure S5, which plots the average number of overlaps per crystal for images simulated at constant D and at constant solids concentration M_T for monodisperse populations of various crystal sizes and aspect ratios. The figure shows that, at a given solids concentration, the average number of overlaps per crystal is a strong function of the particle size and shape while, at a given level of D, the number of overlaps is independent of particle size and shape (notice also the difference in scale of the y-axes). If measurement failure is caused by particle overlap, we would expect the results

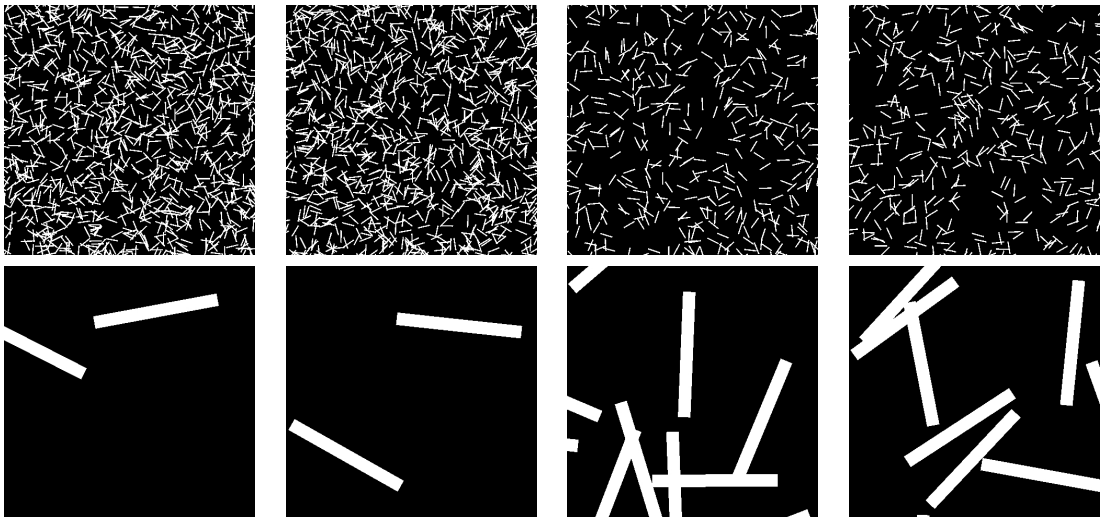


Figure S4: Comparison of images generated for two different monodisperse particle populations at the same solids concentration M_T and at the same D. The top row shows the images for particles with length 1/20th of the image dimension and the bottom row shows images for particles with length one-half the image dimension. The aspect ratio of the particles is 10.

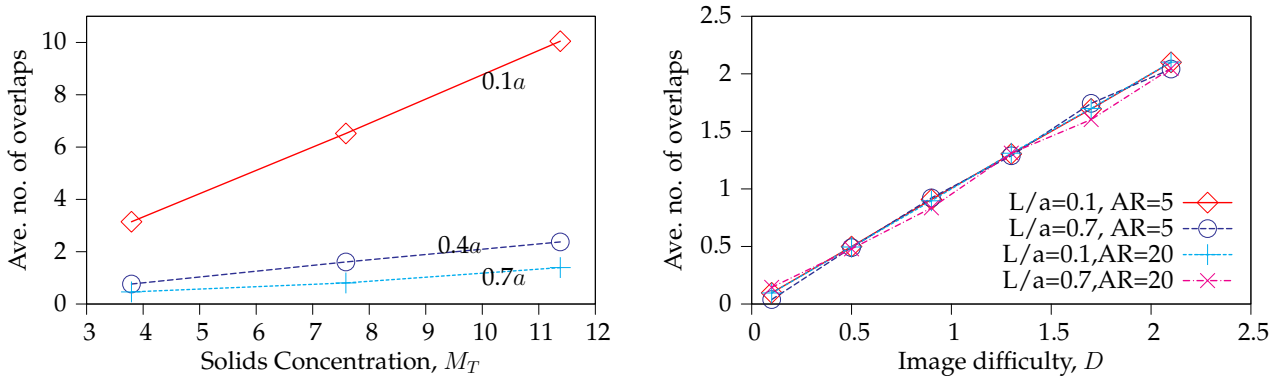


Figure S5: Comparison of average number of overlaps per crystal for images simulated at constant D and at constants solids concentration for monodisperse populations of various crystal sizes.

of automated image analysis at a given D to be relatively independent of the size and shape of the particles. Figure S6 shows the percentage of particles missed by automated image analysis at different levels of solids concentrations and D. As expected, the percent missed is similar for various particle sizes when considered in terms of D, but is vastly different when considered in terms of solids concentration.

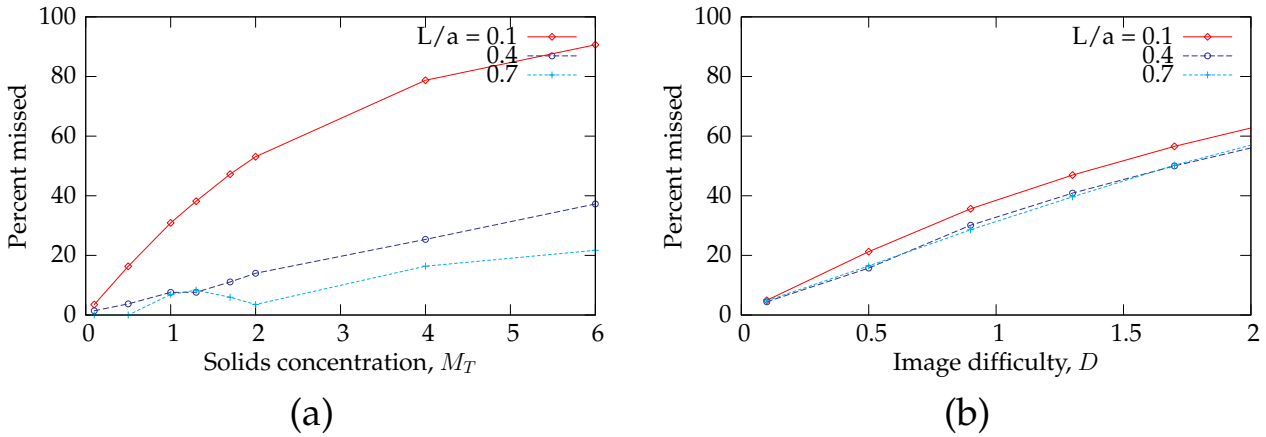


Figure S6: Comparison of percentage of particles missed by automated image analysis for images simulated at constant D and at constants solids concentration for monodisperse populations of various crystal sizes.

8.5.2 Estimation of number density

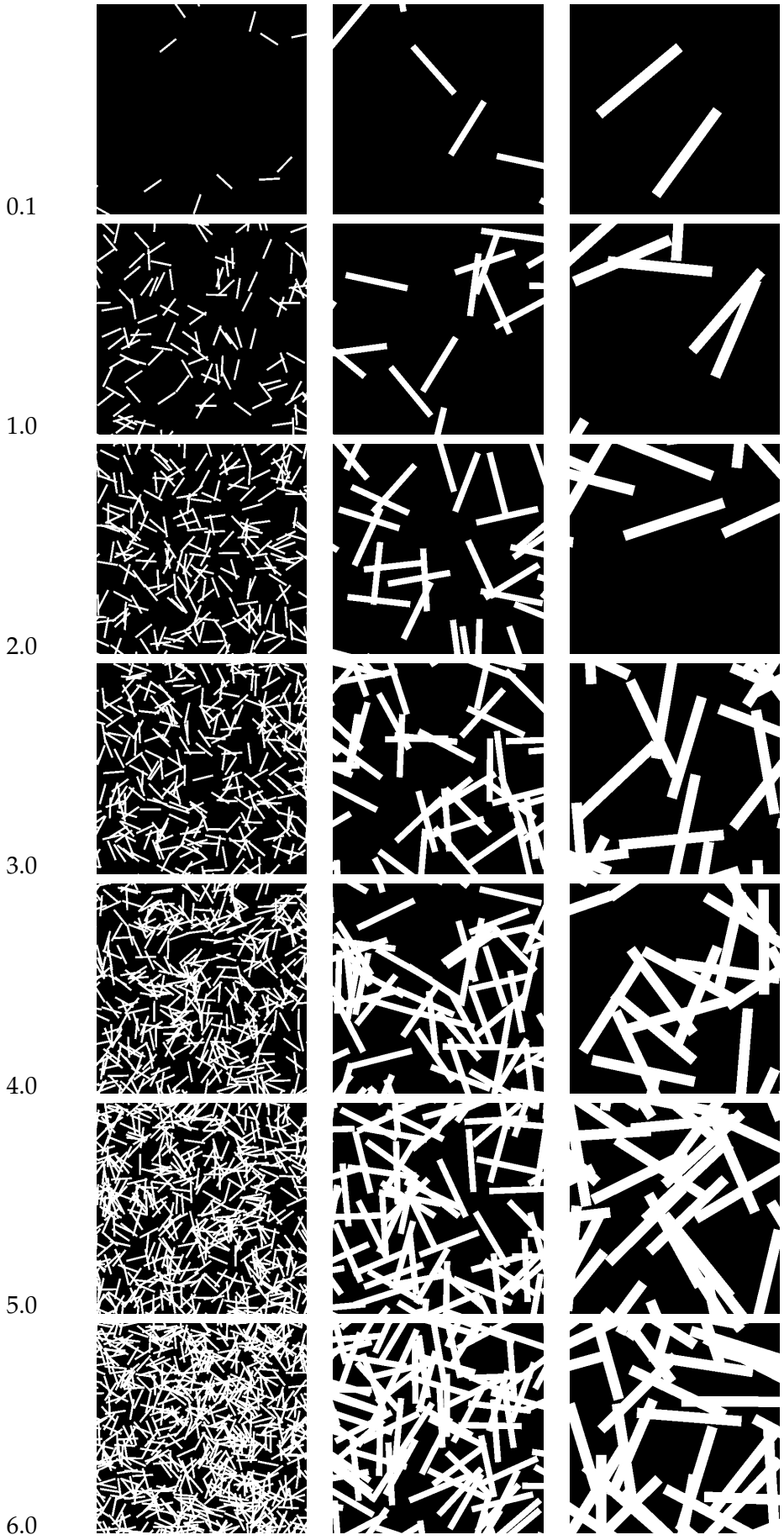
In this section, we examine different methods for estimating the number density ρ. We first examine the behavior of the Miles-Lantuejoul estimator, which corrects for edge effects but not for particle overlap. Next, we examine the performance of the maximum likelihood estimator presented in Section 8.3.4. The estimators are applied to data acquired by analyzing images using SHARC. Example images for various levels of D and three different particle sizes (L/a = 0.1, 0.3, and 0.5) are shown in Figure S5.

Miles-Lantuejoul method

Figure S8 shows the ratio of the estimated and true number density as a function of D for various particle sizes. In this figure, the number density of particles is estimated using the Miles-Lantuejoul method [72, 57], which does not account for overlap. The particle size and shape measurements are obtained using the SHARC algorithm. The estimator's bias increases with D, or as the amount of overlap increases. Given the well-behaved and relatively size-independent correlation between ρ/ρ and D, one may attempt to correct ρ based on an estimated value of D. The result of such an approach is shown in Figure S8(b). The number density estimates for this figure are given by

$$\rho = \frac{\rho_{ML}}{g(D)}$$

in which ρ_{ML} denotes the Miles-Lantuejoul estimate, D is an estimate of D given by $D = \rho_{ML}\Omega$, and g is an empirical function generated by fitting $g(D) = \exp(-\theta D)$ to the data in Figure S8(a). The parameter θ is determined using nonlinear least-squares minimization. This approach is ineffective because D cannot be estimated independently of ρ, so both D and ρ are underestimated.


0.1
1.0
2.0
3.0
4.0
5.0
6.0

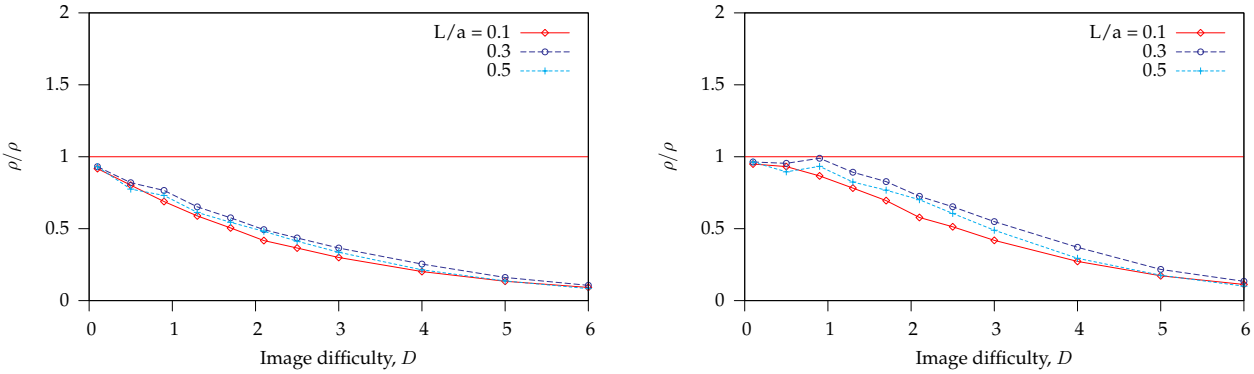


Figure S8: Results of number density estimation using Miles-Lantuejoul method for various particle sizes and various levels of image difficulty. (a) Results for applying Miles-Lantuejoul method directly to image analysis output. (b) Results for empirical correction of the Miles-Lantuejoul estimate based on estimated value of D.

Maximum likelihood method

To implement the maximum likelihood estimator given by Equation (8.8), we first need to complete the probabilistic model by determining the empirical parameter θ. We determine θ by solving the nonlinear least squares minimization problem given by

$$\theta = \arg\min_{\theta} \sum_{i=1}^{n_d} \left(\frac{x_i}{N} - h(\rho_i, \theta) \right)^2 \tag{8.12}$$

$$\text{subject to} \quad 0 \leq \theta \leq 1 \tag{8.13}$$

in which x_i is the number of particles identified by analysis of N images acquired at a given number density ρ_i, and $h(\rho, \theta)$ is the model prediction of the average number of particles identified per image, given by

$$h(\rho, \theta) = \rho N A_{NB} \exp(-\rho \Omega \theta) \tag{8.14}$$

A value of $\theta = 1$ indicates that the image analysis algorithm finds only particles that are completely isolated while a value of $\theta = 0$ indicates the image analysis algorithm identifies every particle in the image, regardless of overlap.

Figure S9 shows the optimal fit of the model in Equation (8.14) to the data observed by applying SHARC to artificial images generated at various number densities. The values of D corresponding to the number densities range from 0.1 to 6.0. Figure S9 indicates that the empirical overlap model gives an excellent prediction of SHARC's behavior for a wide range of densities for $L/a = 0.1$. For $L/a = 0.3$ and $L/a = 0.5$, the model predictions agree with the data only for low densities, corresponding to $D < 3.0$. Accordingly, the maximum likelihood estimator in Equation (8.8) provides good estimates of ρ if $L/a = 0.1$ for all conditions considered, while providing good estimates for $L/a = 0.3$ and $L/a = 0.5$ only for suitably low values of D, as shown in Figure S10.

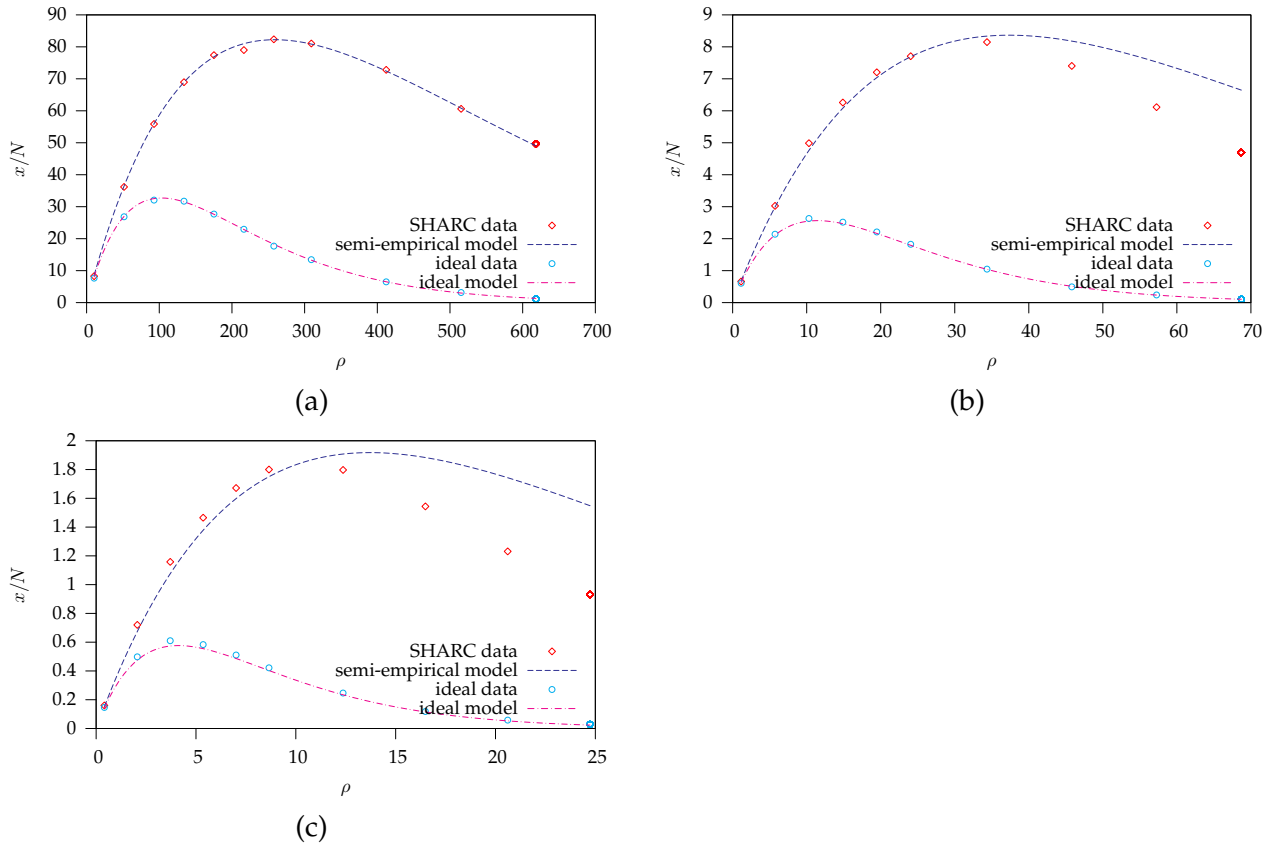


Figure S9: Optimal fit for model of number of particles identified per image by automated image analysis. Also shown is the ideal model, or the expected number of isolated particles predicted for $\theta = 1$ and the number of isolated particles observed in the simulated images (ideal data). (a) $L/a = 0.1$, $\theta = 0.4$; (b) $L/a = 0.3$, $\theta = 0.3$; (c) $L/a = 0.5$, $\theta = 0.3$.

8.5.3 Comments on implementation of maximum likelihood estimator

The likelihood L corresponds to the probability of observing the data given ρ. For the case of $D = 0.1, L/a = 0.1$, the total number of particles appearing in the image is around 10, with almost no overlap. Thus, the total number of particles is essentially equal to the number of isolated particles. However, it is also likely that only 10 isolated particles would be observed if the number density of particles is so high that nearly all particles are overlapping. Thus, in Figure S11, two spikes are observed in the likelihood function: one corresponding to the low concentration case when the total number of particles is close to the number of isolated particles (corresponding to the example image in Figure S11(a)), and one corresponding to the high concentration case (see example image in Figure S11(b)). To determine which maxima corresponds to the true value of ρ, additional information is required that makes it clear whether the system is in the low number density regime or high number density regime. Kellerer [56], for example, incorporates the number of voids (i.e. regions completely surrounded by particles) to calculate the expected number of clumps of

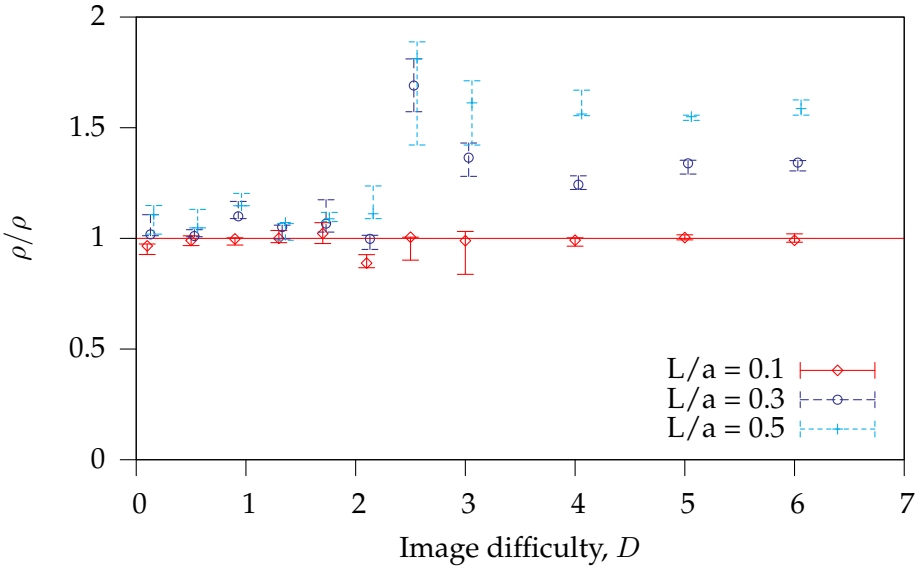


Figure S10: Ratio of estimated density and true density versus image difficulty using SHARC data and empirical correction factors calculated for each different particle size.

particles. The additional information available depends on the type of image analysis algorithm used. The first step of the SHARC algorithm is to identify lines in the image, and we use the number of lines identified by SHARC to determine which of the two peaks to choose. The correct peak is defined as the peak corresponding to a value of ρ that predicts most nearly the number of lines identified by SHARC. A rough prediction of the number of lines is given by $n_l(\rho) = 4\rho A_I$.

8.6 Conclusion

A practical approach has been developed for assessing the reliability of number density estimates obtained using imaging measurements in the presence of particle overlap. The single dimensionless parameter D correlates with the measurement reliability based on the amount of particle overlap. Thus, the parameter D can be used to estimate the errors in the measurements and to aid practitioners in determining the sampling conditions necessary to obtain reliable measurement of particle number density.

It has been shown that the Miles-Lantuejoul estimator, which accounts for edge effects but not particle overlap, underestimates the number density. A maximum likelihood estimator that accounts for both edge effects and particle overlap has been presented. The estimator is based on a semi-empirical model of the probability that a given particle will be correctly identified by automatic image analysis. For a given particle size, a single empirical parameter is sufficient to enable effective number density estimation for a wide range of conditions, particularly for systems in which the particles' dimensions are significantly smaller than the image dimensions.

Various issues should be addressed in future work. Most importantly, an extension of the

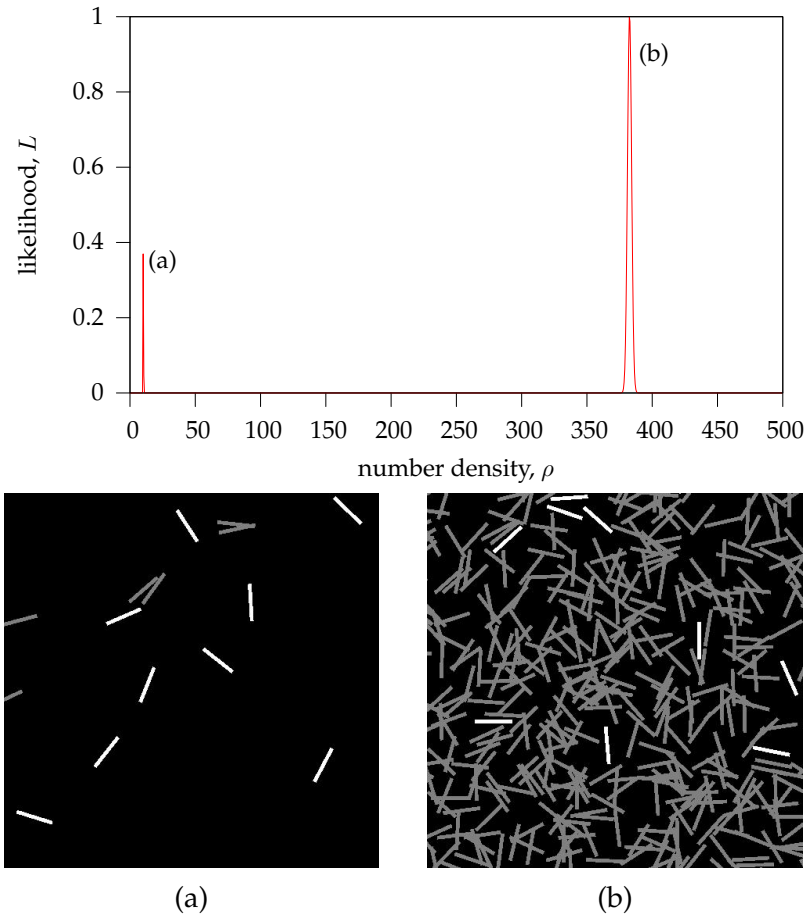


Figure S11: Likelihood of observing n non-overlapping particles with example images at low and high number densities giving the same number of non-overlapping particles. Overlapping particles appear gray while non-overlapping particles appear white.

methods developed in this paper should be developed for polydisperse systems. Armitage [4] and Roach [91, p.46] give discussions that may provide a good starting point for such investigations. The incorporation of these methods into a state estimator for feedback control of number density should also be investigated.

Chapter 9

Parameter Estimation Using Imaging Data *OR* Feasibility of PSD reconstruction in finite time

Chapter 10

Conclusion

10.1 Contributions

10.2 Future work

Appendix A

Derivations for maximum likelihood estimation PSD

A.1 Maximum likelihood estimation of PSD

Let $\mathbf{X}_k = (X_{1k}, \ldots, X_{Tk})$ be a T-dimensional random vector in which X_{ik} gives the number of *non-border* particles of size class i observed in image k. A non-border particle is a particle that is completely enclosed within the imaging volume. A *border* particle, on the other hand, is only partially enclosed within the imaging volume such that only a portion of the particle is observable. For border particles, only the observed length (i.e. the length of the portion of the particle that is inside the imaging volume) can be measured. Accordingly, we let $\mathbf{Y}_k = (Y_{1k}, \ldots, Y_{Tk})$ be a T-dimensional random vector in which Y_{jk} gives the number of *border* particles with observed lengths in size class j that are observed in image k. We denote the observed data, or the realizations of the random vectors $\mathbf{X}_k$ and $\mathbf{Y}_k$, as $\mathbf{x}_k$ and $\mathbf{y}_k$, respectively.

The particle population is represented completely by the vectors $\rho = (\rho_i, \ldots, \rho_T)$ and $\mathbf{S} = (S_1, \ldots, S_{T+1})$ in which ρ_i represents the number of particles of size class i per unit volume and S_i is the lower bound of size class i. Given the data $\mathbf{x}$ and $\mathbf{y}$ (the subscript k denoting the image index is removed here for simplicity), the maximum likelihood estimator of ρ is defined as

$$\rho_b = \arg\max_{\rho} p_{XY}(x_1, y_1, x_2, y_2, \ldots, x_T, y_T|\rho) \tag{A.1}$$

in which in which the subscript b indicates the use of border particle measurements and p_{XY} is the joint probability density for $\mathbf{X}$ and $\mathbf{Y}$. In other words, we want to determine the value of ρ that maximizes the probability of observing exactly x_1 non-border particles of size class 1, y_1 border particles of size class 1, x_2 non-border particles of size class 2, y_2 border particles of size class 2, and so on.

A simplified expression for p_{XY} can be obtained by noting that, at least at low solids concentrations, the observations X_1, Y_1, $\ldots X_T$, Y_T can be assumed to be independent. This assumption means that the observed number of particles of a given size class depends only on the density of particles in that same size class. At high solids concentrations, this assumption seems unreasonable because the number of particle observations in a given size class is reduced due to occlusions by particles in other size classes. At low concentrations, however, the likelihood of occlusion is low. The independence assumption does *not* imply that the observations are not correlated. Rather, the assumption implies that any correlation between observations is due to their dependence on a common set of parameters. As an example, if we observe a large number of non-border

particles, we would expect to also observe a large number of border particles. This correlation can be explained by noting that the probability densities for both border and non-border observations depend on a common parameter, namely, the density of particles. Given the independence assumption, we express p_{XY} as

$$p_{XY} = \prod_{i=1}^{T} p_{X_i}(x_i|\rho) \prod_{j=1}^{T} p_{Y_j}(y_j|\rho) \tag{A.2}$$

in which p_{X_i} and p_{Y_j} are the probability densities for the random variables X_i and Y_j. Using Equation (A.2), the estimator in Equation (A.1) can be reformulated as

$$\rho_b = \arg\min_{\rho} \sum_{i=1}^{T} -\log p_{X_i}(x_i|\rho) - \sum_{j=1}^{T} \log p_{Y_j}(y_j|\rho) \tag{A.3}$$

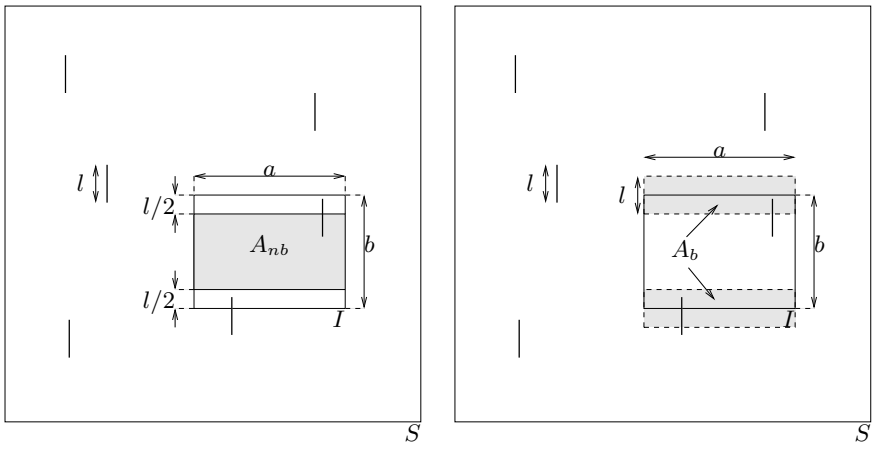
The probability densities p_{X_i} and p_{Y_j} are derived in the following sections. These derivations show that $X_i \sim \text{Poisson}(m_{X_i})$, or that X_i has a Poisson distribution with parameter $m_{X_i} = \rho_i \alpha_i$, in which α_i is a function of the field of view, depth of field, and the lower and upper bounds of size class i. Furthermore, $Y_j \sim \text{Poisson}(m_{Y_j})$, in which $m_{Y_j} = \sum_{i=1}^{T} \rho_i \beta_{ij}$

A.2 Derivation of probability densities

The probability densities p_{X_i} and p_{Y_i} in Equation (A.3) can be derived given the particle geometry and the spatial and orientational probability distributions. Here, we derive p_{X_i} and p_{Y_i} for needle-like particles assuming the particles are randomly, uniformly distributed, both in their 3-dimensional spatial location and in their orientation in the plane perpendicular to the optical axis. To simplify the discussion, we initially present the derivation assuming a 2-dimensional system with monodisperse, vertically-oriented particles. Later, we relax these assumptions and present the derivation for randomly-oriented, polydisperse particles in 3-dimensional space.

A.2.1 Non-border particles

Let S be a square domain in $\mathscr{R}^2$ with dimension B and area $A_S = B^2$. Let I be a rectangular domain in $\mathscr{R}^2$ with horizontal and vertical dimensions a and b, respectively, and area $A_I = ab$. Assume $A_I \ll A_S$ and $I \in S$. Let n_{tot} be the total number of vertically-oriented particles with midpoints randomly and uniformly distributed in S, and define $\rho = n_{\text{tot}}/A_S$ as the density of particles per unit area. Let the length of all particles be l, with $l < \min(a, b)$, and define A_{nb} as the area of the domain in which particles are inside I but do not touch the border of I. Because the particles are oriented vertically, it is easy to show that $A_{nb} = a(b - l)$, as depicted in Figure S1(a). Finally, let X be a random variable denoting the number of non-border particles appearing in I. Assuming the location of each particle in S is independent of the remaining particles' locations, the probability that a *specific* particle will appear entirely within I is given by $p = A_{nb}/A_S$. Given the above assumptions, this probability is constant for all particles. The probability of observing x



(a) (b)

Figure S1: Depiction of hypothetical system of vertically-oriented particles randomly and uniformly distributed in space.

non-border particles in I is analogous to the probability of observing x successes in n_{tot} Bernoulli trials in which the probability of success in each trial is p. Thus, X is a binomial random variable with probability distribution

$$p_X(x) = \begin{pmatrix} n_{\text{tot}} \\ x \end{pmatrix} p^x (1-p)^{n_{\text{tot}}-x}$$

Now, assume $B \to \infty$ while keeping ρ constant. Then $n_{\text{tot}} \to \infty$ and $p = A_{nb}/A_S = A_{nb}\rho/n_{\text{tot}} \to 0$ while $Np = \rho A_{nb}$ remains constant. The limiting distribution of X is therefore Poisson

$$p_X(x) = \frac{e^{-m_X} m_X^x}{x!}, \quad m_X = \rho A_{nb}$$

To extend the analysis to polydisperse, randomly-oriented needles, we discretize the length scale into T size classes and let $\mathbf{X} = (X_1, \ldots, X_T)$ be a T-dimensional random vector in which X_i gives the number of *non-border* particles of size class i observed in a single image. An orientation Θ and length L are assigned to each particle, where $\Theta_1, \Theta_2, \ldots, \Theta_{n_{\text{tot}}}$ are i.i.d. with density function $p_\Theta(\theta), \theta \in [-\pi/2, \pi/2)$ and $L_1, L_2, \ldots, L_{n_{\text{tot}}}$ are i.i.d. with density function $p_L(l), l \in (0, \inf)$. Θ and L are independent of each other and independent of the particle's spatial location. We define $\mathbf{S}$ as the T+1-dimensional vector of breaks between size classes. A particle of length l belongs to size class i if $S_i \leq l < S_{i+1}$. Let $\Delta_i = S_{i+1} - S_i$. Our goal is to determine the probability that a particle of size class i will appear entirely inside the image I, given its density ρ_i. Following the approach used to solve the Buffon-Laplace needle problem [102, p. 4], Figure S2 shows geometrically that for a given orientation θ and length l, $A_{nb}(l, \theta)$ can be calculated as

$$A_{nb}(l,\theta) = \begin{cases} (a - l\cos\theta)(b - l\sin\theta) & 0 \leq \theta \leq \pi/2 \\ (a - l\cos\theta)(b + l\sin\theta) & -\pi/2 \leq \theta \leq 0 \end{cases} \tag{A.4}$$

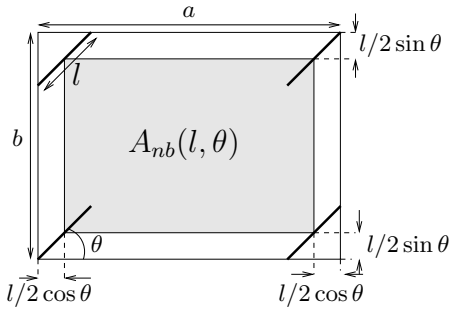


Figure S2: Depiction of geometrical properties used to derive the non-border area function $A_{nb}(l, \theta)$.

The probability that a given particle in size class i will appear entirely within I is given by

$$p_i = \frac{\int_{S_i}^{S_{i+1}} \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} A_{nb}(l,\theta) p_\Theta(\theta) p_L(l) d\theta dl}{\int_{S_i}^{S_{i+1}} \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} A_S p_\Theta(\theta) p_L(l) d\theta dl} \tag{A.5}$$

Thus, the probability that a specific particle of size class i will appear entirely within the image is given by $p_i = \alpha_i / A_S$, where α_i is the numerator in Equation (A.5). Following the same arguments as above, we can show that for an infinitely large system, X_i is a Poisson random variable with parameter $m_{X_i} = \rho_i \alpha_i$.

Extending the analysis to three-dimensional space is trivial because we assume the particles are oriented in the plane perpendicular to the camera's optical axis and assume no interaction between particles. Thus, for a three-dimensional system, X_i is a Poisson random variable with parameter $m_{X_i} = \rho_i \alpha_i$, with $\alpha_i = \alpha_i d$, in which d is the depth of field.

Assuming Θ is distributed uniformly and L is distributed uniformly across each size class, α_i can be calculated as follows. Let $\Delta S_i = S_{i+1} - S_i$, $S_{\max} = \sqrt{a^2 + b^2}$, $\Delta S_{i,\max} = S_{\max} - S_i$, and assume $a > b$. For $S_{i+1} \leq b$,

$$\alpha_i = \frac{d}{\pi \Delta S_i} \left[\frac{1}{3} \left(S_{i+1}^3 - S_i^3 \right) - (a+b) \left(S_{i+1}^2 - S_i^2 \right) + ab\pi \Delta S_i \right]$$

For $b < S_i, S_{i+1} \leq a$,

$$\begin{aligned} \alpha_i = \frac{d}{\pi \Delta S_i} \Bigg[& S_{i+1} \left(a\sqrt{S_{i+1}^2 - b^2} + 2ab \sin^- \left(\frac{b}{S_{i+1}} \right) - b \right) \cdots \\ & - S_i \left(a\sqrt{S_i^2 - b^2} + 2ab \sin^- \left(\frac{b}{S_i} \right) + b \right) \cdots \\ & + ab^2 \log \left(\frac{S_{i+1} + \sqrt{S_{i+1}^2 - b^2}}{S_i + \sqrt{S_i^2 - b^2}} \right) - a \left(S_{i+1}^2 - S_i^2 \right) \Bigg] \end{aligned}$$

For $a \leq S_i, S_{i+1} \leq S_{\text{max}}$,

$$\begin{aligned}\alpha_i = \frac{d}{\pi \Delta S_i} \Bigg\{ & S_{i+1} \left[a\sqrt{S_{i+1}^2 - b^2} + b\sqrt{S_{i+1}^2 - a^2} + 2ab \left(\sin^- \left(\frac{b}{S_{i+1}} \right) - \cos^- \left(\frac{a}{S_{i+1}} \right) \right) \right] \cdots \\ & - S_i \left[a\sqrt{S_i^2 - b^2} + b\sqrt{S_i^2 - a^2} + 2ab \left(\sin^- \left(\frac{b}{S_i} \right) - \cos^- \left(\frac{a}{S_i} \right) \right) \right] \cdots \\ & + ab^2 \log \left(\frac{S_{i+1} + \sqrt{S_{i+1}^2 - b^2}}{S_i + \sqrt{S_i^2 - b^2}} \right) + a^2 b \log \left(\frac{S_{i+1} + \sqrt{S_{i+1}^2 - a^2}}{S_i + \sqrt{S_i^2 - a^2}} \right) \cdots \\ & - (a^2 + b^2) \Delta S_i - \frac{1}{3} \left(S_{i+1}^3 - S_i^3 \right) \Bigg\}\end{aligned}$$

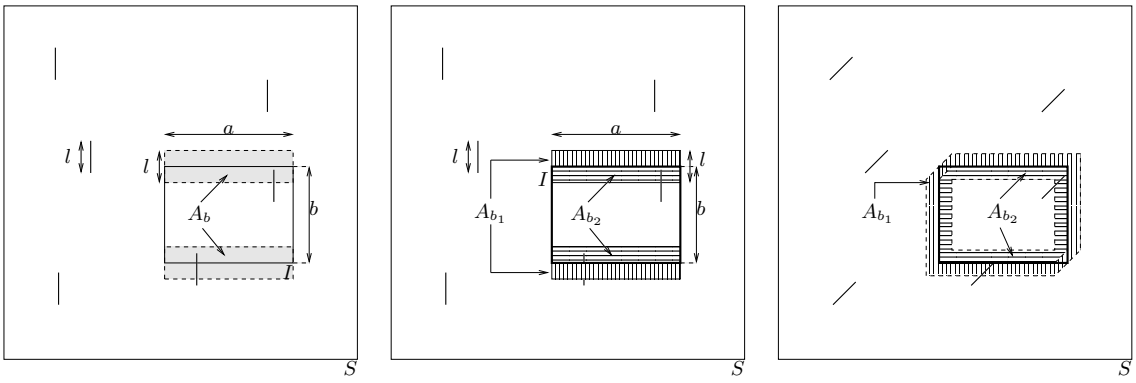
For $a \leq S_i \leq S_{\text{max}}$ and $S_{i+1} > S_{\text{max}}$,

$$\begin{aligned}\alpha_i = \frac{d}{\pi \Delta S_{i,\text{max}}} \Bigg\{ & S_{\text{max}} \left[a\sqrt{S_{\text{max}}^2 - b^2} + b\sqrt{S_{\text{max}}^2 - a^2} + 2ab \left(\sin^- \left(\frac{b}{S_{\text{max}}} \right) - \cos^- \left(\frac{a}{S_{\text{max}}} \right) \right) \right] \cdots \\ & - S_i \left[a\sqrt{S_i^2 - b^2} + b\sqrt{S_i^2 - a^2} + 2ab \left(\sin^- \left(\frac{b}{S_i} \right) - \cos^- \left(\frac{a}{S_i} \right) \right) \right] \cdots \\ & + ab^2 \log \left(\frac{S_{\text{max}} + \sqrt{S_{\text{max}}^2 - b^2}}{S_i + \sqrt{S_i^2 - b^2}} \right) + a^2 b \log \left(\frac{S_{\text{max}} + \sqrt{S_{\text{max}}^2 - a^2}}{S_i + \sqrt{S_i^2 - a^2}} \right) \cdots \\ & - (a^2 + b^2) \Delta S_{i,\text{max}} - \frac{1}{3} \left(S_{\text{max}}^3 - S_i^3 \right) \Bigg\}\end{aligned}$$

A.2.2 Border particles

As before, we simplify the discussion by first presenting the derivation of p_{Y_i} for monodisperse, vertically-oriented particles. Let Y be a random variable denoting the total number of border particles appearing in I. Define A_b as the area of the domain in which particles touch the border of I, as depicted in Figure S3(a). For the present system, $A_b = 2al$. The probability that a *specific* particle will touch the border of I is given by $p = A_b/A_S$. Following the same arguments as above, we can show that for an infinitely large system, Y is a Poisson random variable with parameter $m_Y = \rho A_b$.

Now, assume we would like to incorporate additional information into our estimation by taking into account not only the *number* of border particles, but also their observed lengths. For a monodisperse population, these observed lengths can take on values anywhere between 0 and l. We therefore discretize the length scale on [0 l] and let j denote the size class corresponding to the observed length. We define Y_j as a random variable denoting the number of border particles appearing in I with observed length in size class j. Figure S3(b) illustrates this approach for two size classes. In this figure, A_{b_1} is the area of the region in which particles produce observed lengths from 0 to $l/2$, corresponding to size class 1, while A_{b_2} is the area of the region in which particles produce observed lengths from $l/2$ to l, corresponding to size class 2. The probability



(a) (b) (c)

Figure S3: Depiction of hypothetical system of vertically-oriented particles randomly and uniformly distributed in space.

that a specific particle will touch the border of I and produce an observed length in size class j is $p = A_{b_j}/A_S$. Thus, Y_j is a Poisson random variable with parameter $m_{Y_j} = \rho A_{b_j}$.

In Figure S3(b), $A_{b_1} = A_{b_2}$. This equality between the areas of different observed length size classes does not hold in general, however, as illustrated in Figure S3(c). In this figure, we assume all particles are oriented diagonally, at 45 degrees from the horizontal, and the figure illustrates that $A_{b_1} > A_{b_2}$. Hence, in general, border particles are more likely to result in observed lengths in the lower size classes.

To extend the analysis to polydisperse systems with random orientation, we define a new random variable Y_{ij} that gives the number of particles in size class i that intersect the image border, producing an observed length in size class j. Given that the size class of each border particle is unknown, we define the random variable Y_j as the total number of border particles producing observed lengths in size class j, noting that $Y_j = \sum_i Y_{ij}$. Our approach is to determine the probability density for Y_{ij} for all i and to use these densities to derive the probability density for Y_j.

We define the function $A_{b_j}(l, \theta)$ as the area of the region in which a particle of length l and orientation θ produces an observed length corresponding to size class j. To calculate $A_{b_j}(l, \theta)$, it is convenient to define an area function $A(l, \theta, l)$ as the area of the region in which particles of length l and orientation θ either intersect or are enclosed within the image boundary and produce an observed length greater than or equal to l. $A(l, \theta, l)$ can be calculated using the geometric relationships shown in Figure S4: In this figure, the thick-lined, outer rectangle is the image region, and the inner rectangle is the region inside which a particle with length l and orientation θ will be entirely enclosed within the image boundary, thus producing an observed length of exactly l. A particle with its midpoint along the perimeter of the outermost hexagon would touch the image boundary but give an observed length of 0. A particle with its midpoint anywhere inside the innermost hexagon will produce an observed length greater than or equal to l. Using the

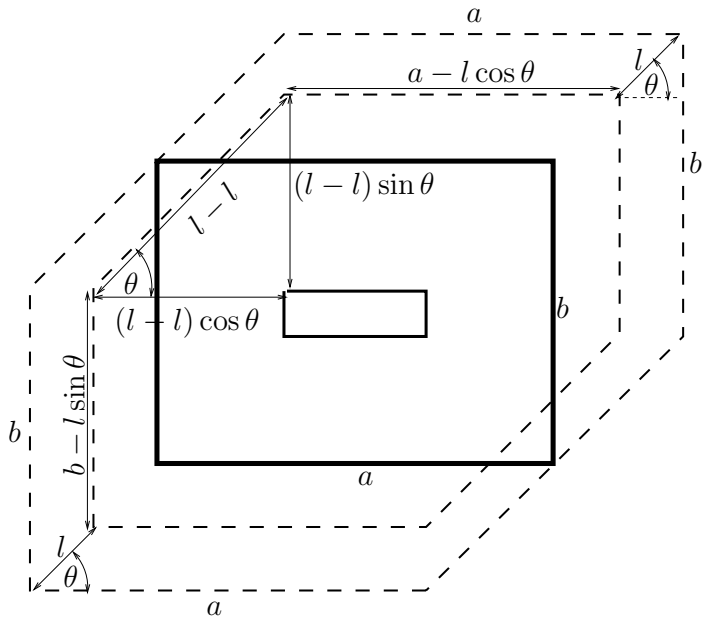


Figure S4: Depiction of non-border area for arbitrary length and orientation.

relationships indicated in this figure, and assuming $l \leq l$, $A(l, \theta, l)$ can be calculated as

$$A(l,\theta,l) = \begin{cases} (a + (l-2l)\cos\theta)(b + (l-2l)\sin\theta) - (l-l)^2 \sin\theta\cos\theta & 0 \leq \theta \leq \pi/2 \\ (a + (l-2l)\cos\theta)(b - (l-2l)\sin\theta) + (l-l)^2 \sin\theta\cos\theta & -\pi/2 \leq \theta \leq 0 \end{cases} \tag{A.6}$$

If $b \leq l < a$, Equation (A.6) is valid only for θ on $(-sin^{-}(b/l), sin^{-}(b/l))$. If $a, b \leq l < (a^2 + b^2)^{1/2}$, Equation (A.6) is valid only for θ on $(-sin^{-}(b/l), -cos^{-}(a/l))$ and $(cos^{-}(a/l), sin^{-}(b/l))$. $A_{b_j}(l, \theta)$ is given by

$$A_{b_j}(l,\theta) = \begin{cases} A(l,\theta,S_j) - A(l,\theta,S_{j+1}) & l \geq S_{j+1} \\ A(l,\theta,S_j) - A_{nb}(l,\theta) & S_j \leq l < S_{j+1} \\ 0 & l < S_j \end{cases} \tag{A.7}$$

The probability that a given particle in size class i will appear within I and produce an observed length in size class j is given by

$$p_{ij} = \frac{\int_{S_i}^{S_{i+1}} \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} A_{b_j}(l,\theta) p_\Theta(\theta) p_L(l) d\theta dl}{\int_{S_i}^{S_{i+1}} \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} A_S p_\Theta(\theta) p_L(l) d\theta dl} \tag{A.8}$$

The probability that a specific particle in size class i will touch the border of I and produce an observed length in size class j is $p_{ij} = \beta_{ij}/A_S$, with β_{ij} being the numerator in Equation (A.8). Thus, for an infinitely large system, Y_{ij} is a Poisson random variable with parameter $m_{Y_{ij}} = \rho_i \beta_{ij}$. Assuming $Y_{1j}, Y_{2j}, \ldots, Y_{Tj}$ are independent, then $Y_j = \sum_i Y_{ij}$ is also a Poisson random variable

with parameter $m_{Y_j} = \sum_i \rho_i \beta_{ij}$ [12, p.440]. As in the non-border case, the analysis is extended to three-dimensional space assuming the particles are oriented in the plane perpendicular to the camera's optical axis and that the particles do not interact. Thus, for a three-dimensional system, Y_j is a Poisson random variable with parameter $m_{Y_j} = \sum_i \rho_i \beta_{ij}$, with $\beta_{ij} = \beta_{ij} d$.

Assuming Θ is distributed uniformly and L is distributed uniformly across each size class, β_{ij} is calculated as follows. Let the length scale discretization be the same for both border and non-border particles. As before, let $\Delta S_i = S_{i+1} - S_i$, $S_{\max} = \sqrt{a^2 + b^2}$, $\Delta S_{i,\max} = S_{\max} - S_i$, and assume $a > b$. Then β_{ij} is given by

$$\beta_{ij} = \begin{cases} A(i, S_j) - A(i, S_{j+1}) & i > j \\ A(i, S_j) - \alpha_i & i = j \\ 0 & i < j \end{cases} \tag{A.9}$$

in which $A(i, S)$ is calculated as

$$A(i,S) = \frac{d}{\pi \Delta S_i} \left[\Delta S_i \left(2ab\gamma_1 - 4bS\gamma_2 - 4aS\gamma_3 + 3S^2\gamma_4 \right) + \left(S_{i+1}^2 - S_i^2 \right) \left(b\gamma_2 + a\gamma_3 - S\gamma_4 \right) \right]$$

$$\gamma_1 = \begin{cases} \pi/2 & S < b \\ \sin^- (b/S) & b < S < a \\ \sin^- (b/S) - \cos^- (a/S) & a < S < S_{\max} \\ \sin^- (b/S_{\max}) - \cos^- (a/S_{\max}) & S > S_{\max} \end{cases}$$

$$\gamma_2 = \begin{cases} 1 & S < b \\ b/S & b < S < a \\ \left(b - \sqrt{S^2 - a^2} \right) / S & a < S < S_{\max} \\ \left(b - \sqrt{S_{\max}^2 - a^2} \right) / S_{\max} & S > S_{\max} \end{cases}$$

$$\gamma_3 = \begin{cases} 1 & S < b \\ 1 - \sqrt{S^2 - b^2}/S & b < S < a \\ \left(a - \sqrt{S^2 - b^2} \right) / S & a < S < S_{\max} \\ \left(a - \sqrt{S_{\max}^2 - b^2} \right) / S_{\max} & S > S_{\max} \end{cases}$$

$$\gamma_4 = \begin{cases} 1 & S < b \\ b^2/S^2 & b < S < a \\ (a^2 + b^2 - S^2)/S^2 & a < S < S_{\max} \\ (a^2 + b^2 - S_{\max}^2)/S_{\max}^2 & S > S_{\max} \end{cases}$$

A.3 Validation of Marginal Densities

To ensure the correctness of the probability densities derived in the previous section, four different Monte Carlo simulations were carried out in which artificial images of particulate populations were generated. Figure S5 shows example images generated for each simulation. Each of these

images has a horizontal image dimension of a=480 pixels and a vertical dimension of b=480 pixels. The first row displays four simulated images for monodisperse particles of length 0.5a with N_c=25 crystals per image. The second row shows images of particles uniformly distributed on [0.1a 0.9a] with N_c=25. The third row shows images of particles normally-distributed with $\mu = 0.5a$ and $\sigma = 0.4a/3$ with N_c=25, and the fourth row shows example images for simulations of particles uniformly-distributed on [0.1a 2.0a] with N_c=15. For each simulation, 20,000 artificial images were generated. Based on the observations in these 20,000 images, a histogram was generated for each size class giving the frequency of observations for both border and non-border particles. These histograms are compared with the theoretical marginal densities in Figures S6– S13.

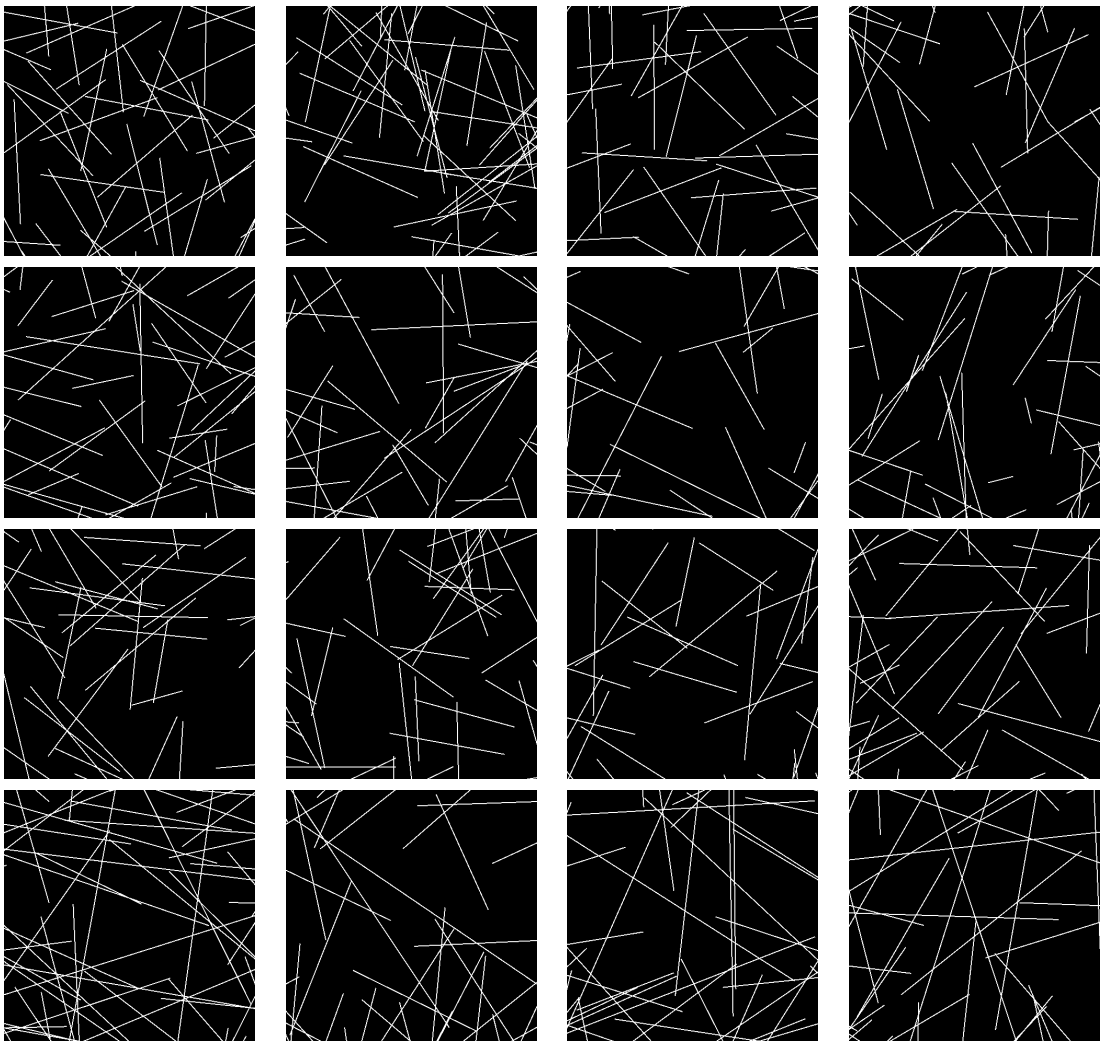


Figure S5: Example images for simulations of various particle populations. Row 1: monodisperse particles of length $0.5a$, N_c=25. Row 2: particles uniformly distributed on [$0.1a$ $0.9a$]. Row 3: particles normally-distributed with $\mu = 0.5a$ and $\sigma = 0.4a/3$ with N_c=25. Row 4: particles uniformly-distributed on [$0.1a$ $2.0a$], N_c=15.

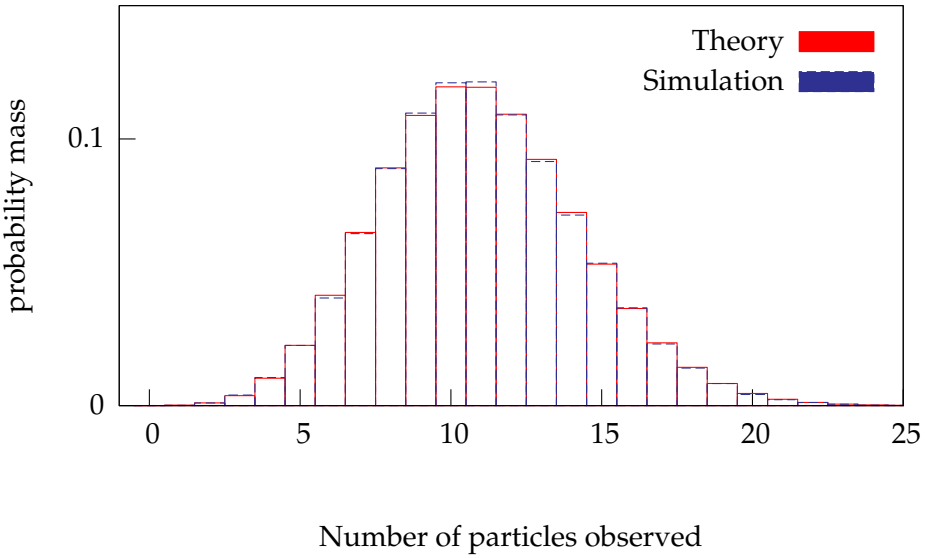


Figure S6: Comparison of theoretical and simulated marginal densities for randomly-oriented, monodisperse particles of length 0.5 and measured by partitioning [0.1 0.9] into ten bins. Results are for non-border particles.

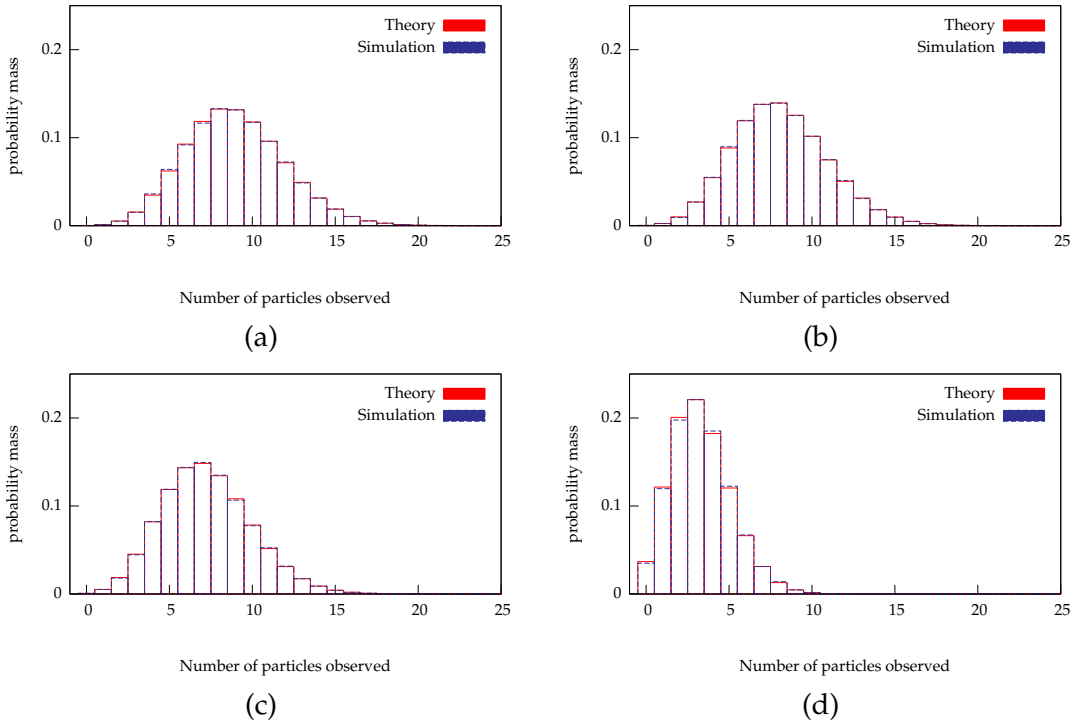


Figure S7: Comparison of theoretical and simulated marginal densities for randomly-oriented, monodisperse particles of length 0.5 and measured by partitioning [0.1 0.9] into ten bins. Results are for border particles.

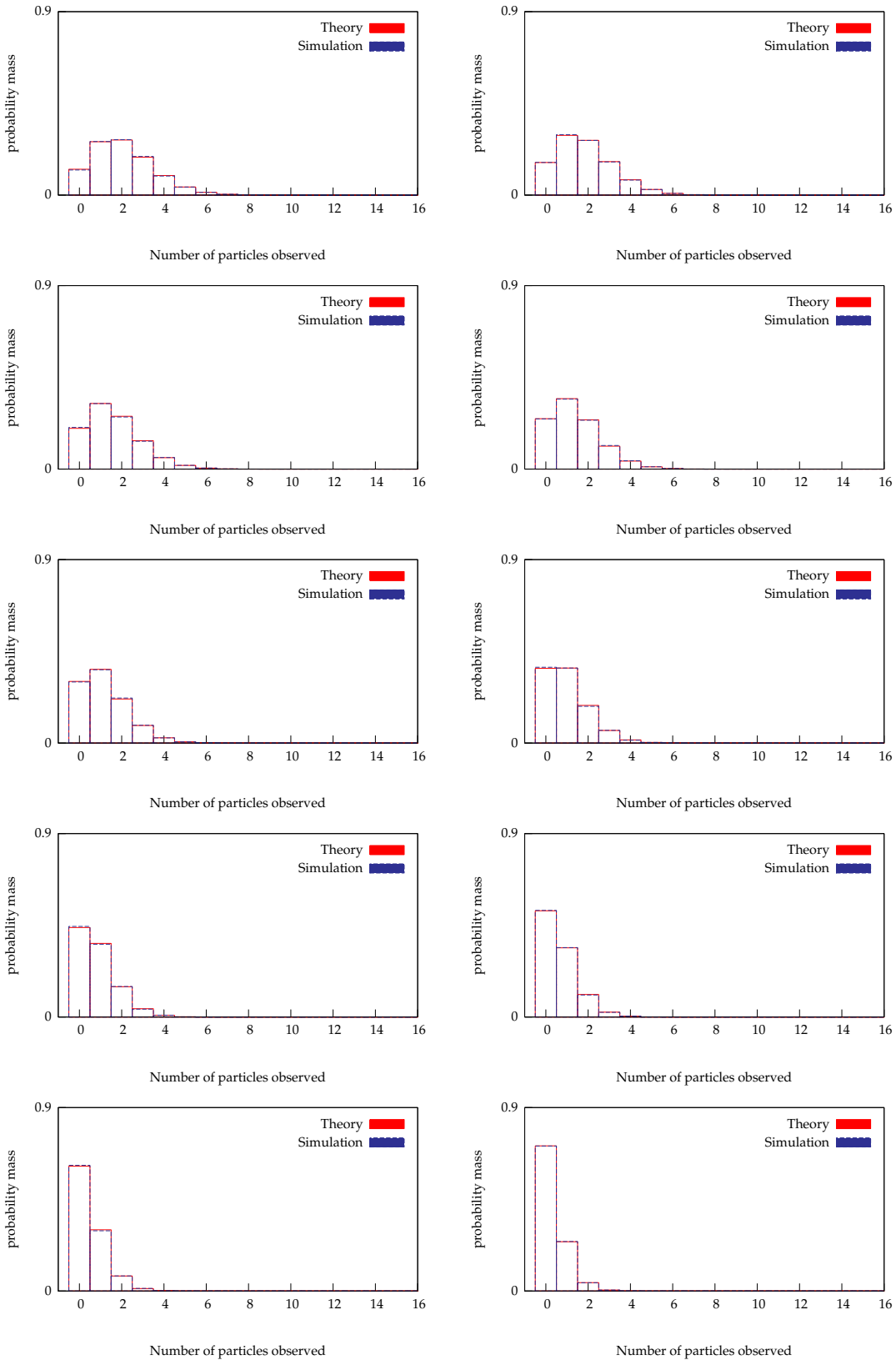


Figure S8: Comparison of theoretical and simulated marginal densities for randomly-oriented particles distributed uniformly on [0.1 0.9] and measured by partitioning [0.1 0.9] into ten bins. Results are for non-border particles.

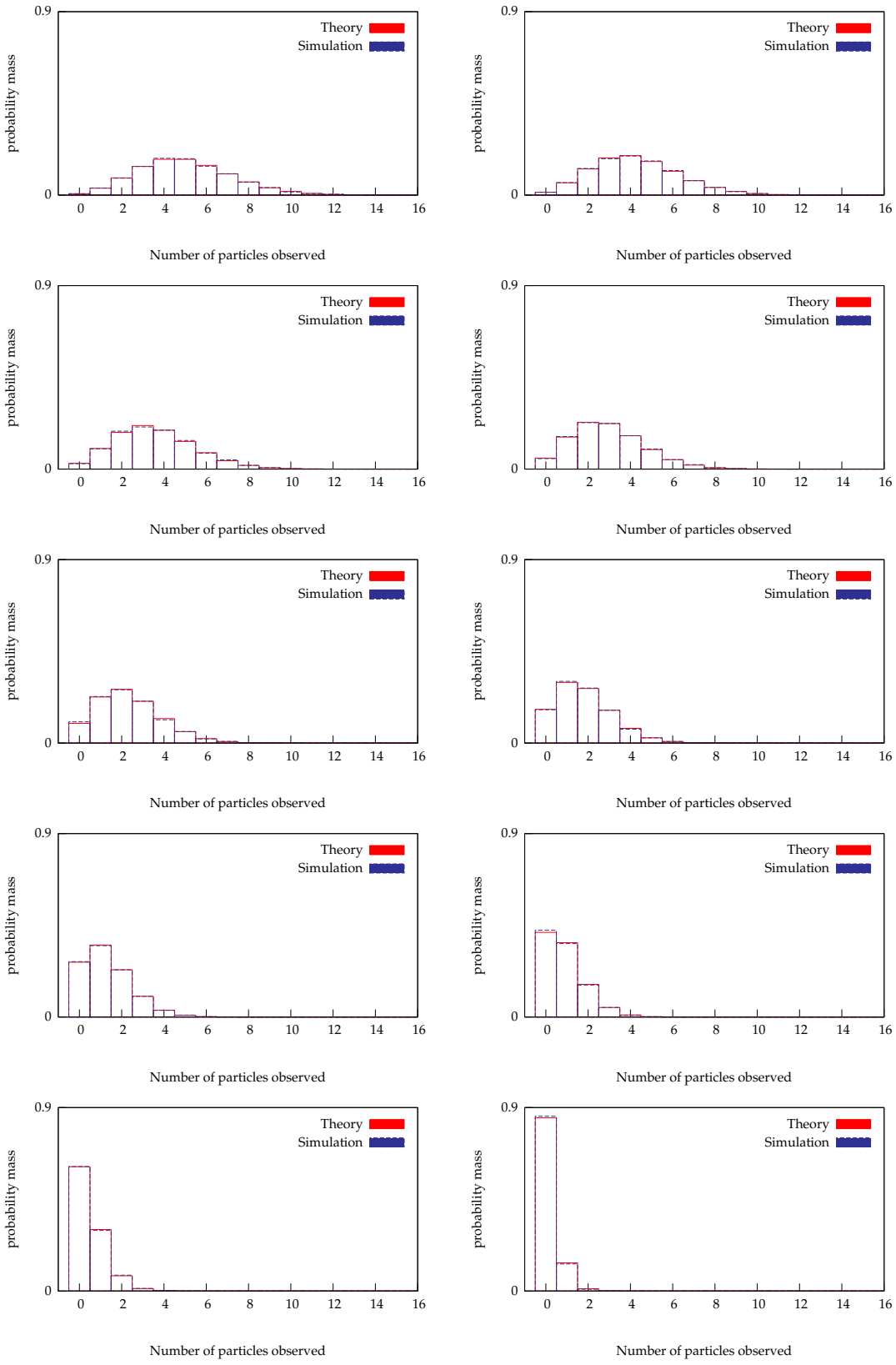


Figure S9: Comparison of theoretical and simulated marginal densities for randomly-oriented particles distributed uniformly on [0.1 0.9] and measured by partitioning [0.1 0.9] into ten bins. Results are for border particles.

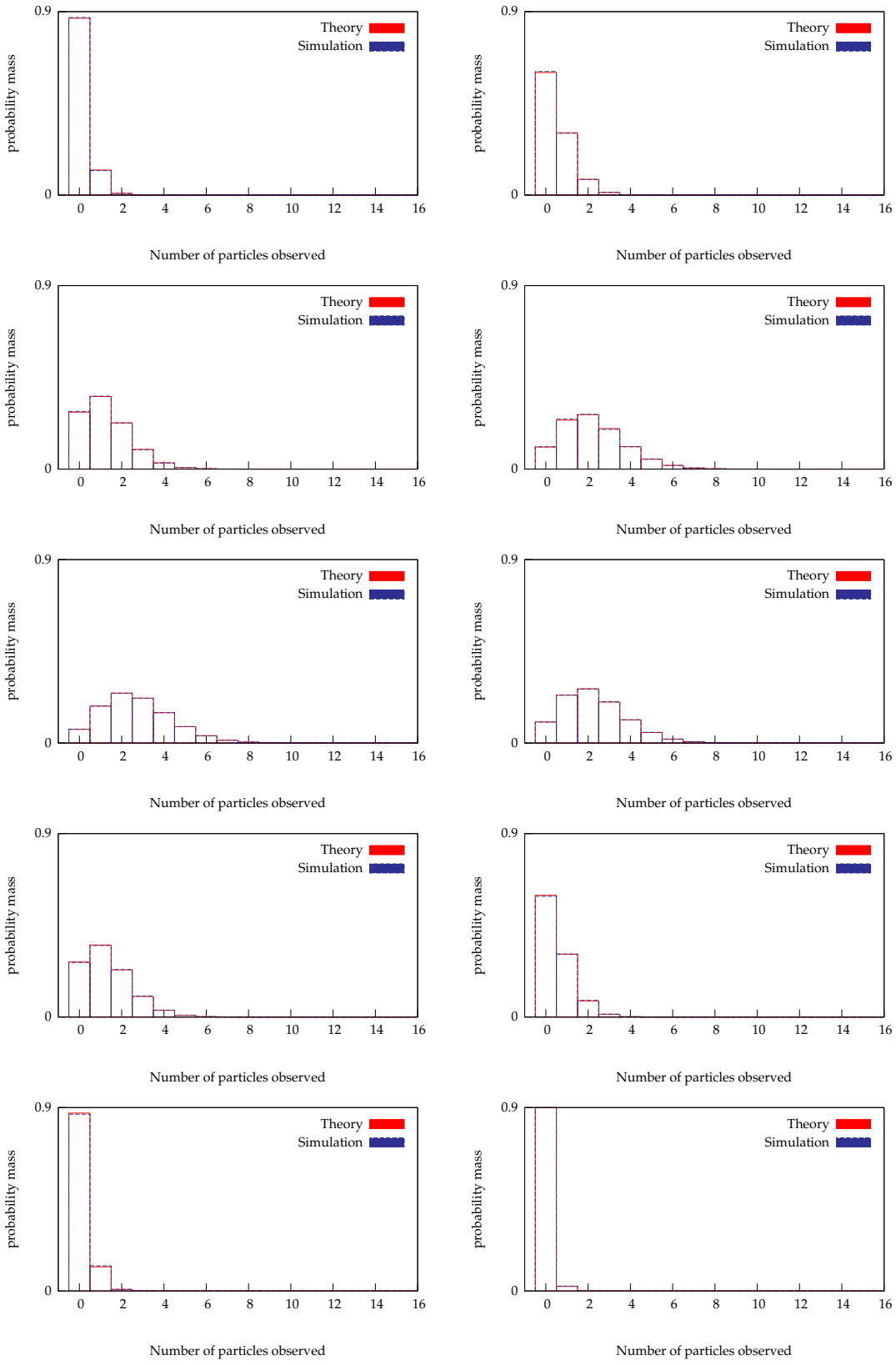


Figure S10: Comparison of theoretical and simulated marginal densities for randomly-oriented particles distributed normally and measured by partitioning [0.1 0.9] into 10 bins. Results are for non-border particles.

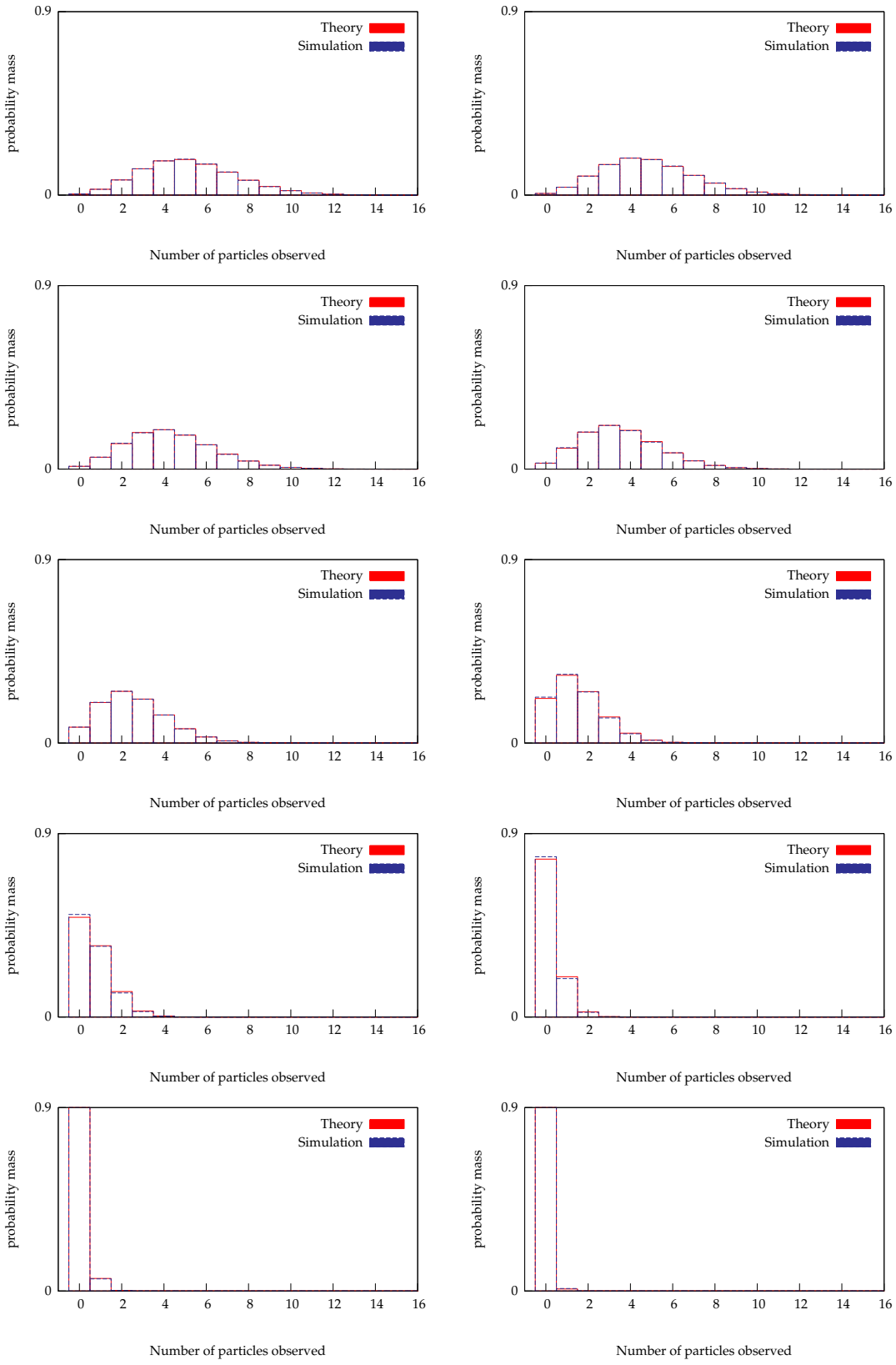


Figure S11: Comparison of theoretical and simulated marginal densities for randomly-oriented particles distributed normally and measured by partitioning [0.1 0.9] into 10 bins. Results are for border particles.

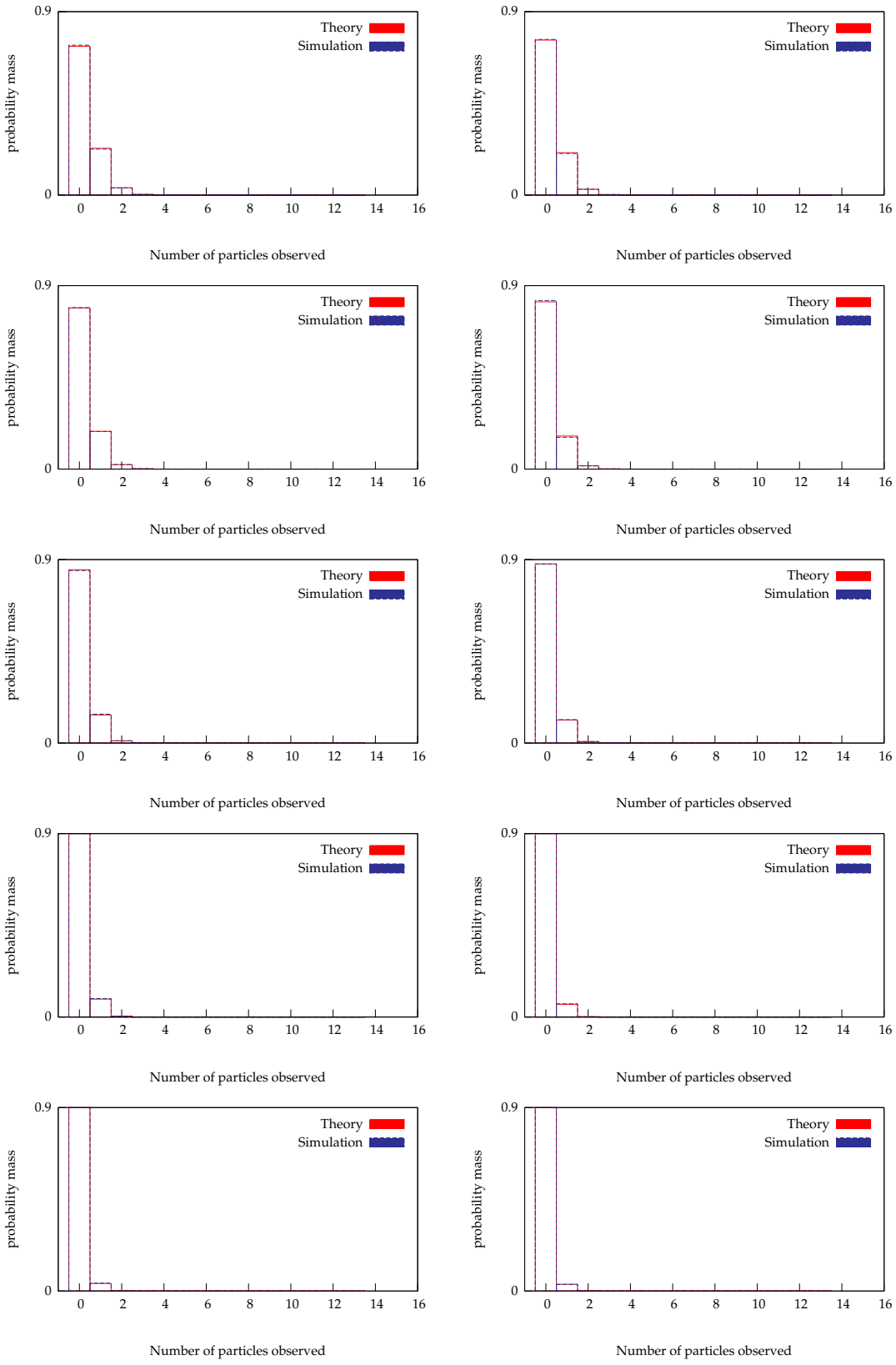


Figure S12: Comparison of theoretical and simulated marginal densities for randomly-oriented particles distributed uniformly on [0.4 2.0] and measured by partitioning [0.4 1.0] into 9 bins with a 10th bin spanning [1.0 $\sqrt{2}$]. Results are for non-border particles.

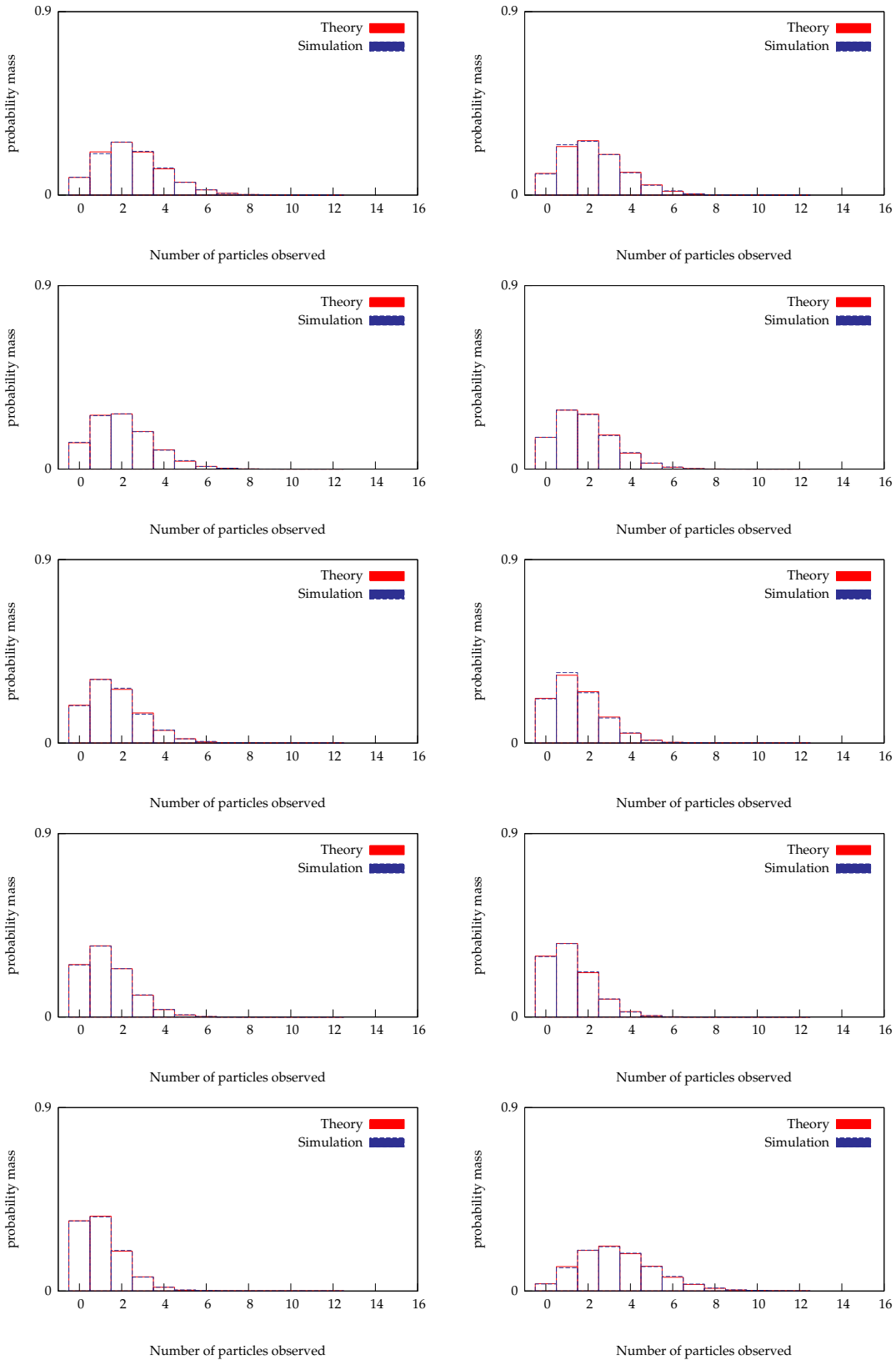


Figure S13: Comparison of theoretical and simulated marginal densities for randomly-oriented particles distributed uniformly on [0.4 2.0] and measured by partitioning [0.4 1.0] into 9 bins with a 10th bin spanning [1.0 $\sqrt{2}$]. Results are for border particles.

Bibliography

[1] *Stochastic Geometry and Its Applications*. John Wiley & Sons, Chichester, 1987.

[2] P. Agarwal and K. A. Berglund. In situ monitoring of calcium carbonate polymorphs during batch crystallization in the presence of polymeric additives using Raman spectroscopy. *Crystal Growth and Design*, 3(6):941–946, 2003.

[3] T. Allen. *Powder Sampling and Particle Size Determination*. Elsevier, 2003.

[4] P. Armitage. An overlap problem arising in particle counting. *Biometrika*, 36(3/4):257–266, December 1949.

[5] A. J. Baddeley. *Stochastic Geometry Likelihood and Computation*, chapter Spatial sampling and censoring, pages 37–78. Chapman and Hall/CRC, Boca Raton, FL, 1999.

[6] A. J. Baddeley. *Stochastic Geometry Likelihood and Computation*, chapter A crash course in stochastic geometry, pages 1–35. Chapman and Hall/CRC, Boca Raton, FL, 1999.

[7] P. Barrett. Selecting in-process particle-size analyzers. *Chemical Engineering Progress*, 99(8):26–32, August 2003.

[8] P. Barrett and B. Glennon. In-line FBRM monitoring of particle size in dilute agitated suspensions. *Particle and Particle Systems Characterization*, 16(5):207–211, 1999.

[9] P. Barrett and B. Glennon. Characterizing the metastable zone width and solubility curve using Lasentec FBRM and PVM. *Chemical Engineering Research and Design*, 80(A7):799–805, 2002.

[10] J. Bauer, S. Spanton, R. Henry, J. Quick, W. Dziki, W. Porter, and J. Morris. Ritonavir: An extraordinary example of conformational polymorphism. *Pharmaceutical Research*, 18(6):859–866, 2001.

[11] M. Birch, S. J. Fussell, P. D. Higginson, N. McDowall, and I. Marziano. Towards a PAT-based strategy for crystallization development. *Organic Process Research & Development*, 9(3):360–364, 2005.

[12] Y. M. M. Bishop, S. E. Fienberg, and P. W. Holland. *Discrete Multivariate Analysis: Theory and Practice*. The MIT Press, Cambridge, Massachusetts, 1975.

[13] N. Blagden, R. Davey, R. Rowe, and R. Roberts. Disappearing polymorphs and the role of reaction by-products: The case of sulphathiazole. *International Journal of Pharmaceutics*, 172(1–2):169–177, 1998.

[14] S. Boerrigter, G. Josten, J. van de Streek, F. Hollander, J.Los, H. Cuppen, P. Bennema, and H. Meekes. MONTY: Monte Carlo crystal growth on any crystal structure in any crystallographic orientation; application to fats. *Journal of Physical Chemistry A*, 108(27):5894–5902, 2004.

[15] R. D. Braatz. Advanced control of crystallization processes. *Annual Reviews in Control*, 26:87–99, 2002.

[16] R. D. Braatz and S. Hasebe. Particle size and shape control in crystallization processes. In *Chemical Process Control—CPC 6*, pages 307–327, Tucson, Arizona, January 2001.

[17] J. B. Burns, A. R. Hanson, and E. M. Riseman. Extracting straight lines. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8(4):425–455, July 1986.

[18] J. Calderon De Anda, X. Wang, X. Lai, and K. Roberts. Classifying organic crystals via in-process image analysis and the use of monitoring charts to follow polymorphic and morphological changes. *Journal of Process Control*, 15(7):785–797, 2005.

[19] J. Calderon De Anda, X. Wang, and K. Roberts. Multi-scale segmentation image analysis for the in-process monitoring of particle shape with batch crystallizers. *Chemical Engineering Science*, 60:1053–1065, 2005.

[20] S. Christy and R. Horaud. Iterative pose computation from line correspondences. *Computer Vision and Image Understanding*, 73(1):137–144, January 1999.

[21] S. H. Chung, D. L. Ma, and R. D. Braatz. Optimal seeding in batch crystallization. *The Canadian Journal of Chemical Engineering*, 77(3):590–596, 1999.

[22] G. Clydesdale, R. Docherty, and K. J. Roberts. HABIT—a program for predicting the morphology of molecular crystals. *Computer Physics Communications*, 64(2):311–328, 1991.

[23] W. W. Daniel. *Applied Nonparametric Statistics*. Houghton Mifflin Company, Boston, MA, 1978.

[24] S. Datta and D. J. W. Grant. Crystal structure of drugs: Advances in determination, prediction and engineering. *Nature Reviews Drug Discovery*, 3(1):42–57, January 2004.

[25] C. Deeley, R. Spragg, and T. Threlfall. A comparison of Fourier-Transform Infrared and Near-Infrared Fourier-Transform Raman-spectroscopy for quantitative measurements — An application in polymorphism. *Spectrochimica Acta A*, 47(9–10):1217–1223, 1991.

[26] M. Dhome, M. Richetin, J.-T. Lapreste, and G. Rives. Determination of the attitude of 3-D objects from a single perspective view. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(12):1265–1278, December 1989.

[27] N. Doki, H. Seki, K. Takano, H. Asatani, M. Yokota, and N. Kubota. Process control of seeded batch cooling crystallization of the metastable α-form glycine using an in-situ ATR-FTIR spectrometer and an in-situ FBRM particle counter. *Crystal Growth and Design*, 4(5):949–953, 2004.

[28] W. Doyle. Technical note AN-923. A user's guide to: spectroscopic analysis for industrial applications. Technical report, Axiom Analytical, Inc., Irvine, CA, January 2005.

[29] D. D. Dunuwila and K. A. Berglund. ATR FTIR spectroscopy for in situ measurement of supersaturation. *Journal of Crystal Growth*, 179(1–2):185–193, 1997.

[30] D. D. Dunuwila, L. B. Carroll II, and K. A. Berglund. An investigation of the applicability of attenuated total reflection infrared spectroscopy for measurement of solubility and supersaturation of aqueous citric acid solutions. *Journal of Crystal Growth*, 137(3–4):561–568, 1994.

[31] A. Etemadi, J.-P. Schmidt, G. Matas, J. Illingworth, and J. Kittler. Low-level grouping of straight line segments. In *Proceedings of the British Machine Vision Conference*, 1991.

[32] J. A. Falcon and K. A. Berglund. Monitoring of antisolvent addition crystallization with Raman spectroscopy. *Crystal Growth and Design*, 3(6):947–952, 2003.

[33] L. Feng and K. A. Berglund. ATR-FTIR for determining optimal cooling curves for batch crystallization of succinic acid. *Crystal Growth and Design*, 2(5):449–452, 2002.

[34] G. Fevotte. New perspectives for the on-line monitoring of pharmaceutical crystallization processes using in situ infrared spectroscopy. *International Journal of Pharmaceutics*, 241(2):263–278, 2002.

[35] G. Fevotte, J. Calas, F. Puel, and C. Hoff. Applications of NIR spectroscopy to monitoring and analyzing the solid state during industrial crystallization processes. *International Journal of Pharmaceutics*, 273(1–2):159–169, 2004.

[36] B. A. Finlayson. *Nonlinear Analysis in Chemical Engineering*. McGraw-Hill, New York, 1980.

[37] D. A. Forsyth and J. Ponce. *Computer Vision: A Modern Approach*. Prentice Hall series in artificial intelligence. Prentice Hall, New Jersey, 2003.

[38] M. Fujiwara, P. S. Chow, D. L. Ma, and R. D. Braatz. Paracetamol crystallization using laser backscattering and ATR-FTIR spectroscopy: Metastability, agglomeration, and control. *Crystal Growth and Design*, 2(5):363–370, 2002.

[39] L. Gardini, A. Servida, M. Morbidelli, and S. Carra. Use of orthogonal collocation on finite elements with moving boundaries for fixed bed catalytic reactor simulation. *Computers & Chemical Engineering*, 9(1):1–17, 1985.

[40] G. H. Givens and J. A. Hoeting. *Computational Statistics*. Wiley Series in Probability and Statistics. John Wiley & Sons, New Jersey, 2005.

[41] R. C. Gonzalez and R. E. Woods. *Digital Image Processing*. Prentice Hall, second edition, 2002.

[42] D. A. Green. Particle formation and morphology control. *DuPont Magazine*, College Report supplement, November/December 1988.

[43] W. Grimson and D. Huttenlocher. On the verification of hypothesized matches in model-based recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(12):1201–1213, 1989.

[44] H. Gron, A. Borissova, and K. Roberts. In-process ATR-FTIR spectroscopy for closed-loop supersaturation control of a batch crystallizer producing monosodium glutamate crystals of defined size. *Industrial and Engineering Chemistry Research*, 42(1):198–206, 2003.

[45] P. Hall. Correcting segment counts for edge effects when estimating intensity. *Biometrika*, 72(2):459–63, 1985.

[46] R. W. Hartel. Crystallization in foods. In A. S. Myerson, editor, *Handbook of industrial crystallization*, pages 287–304. Butterworth-Heinemann, USA, 2nd edition, 2002.

[47] M. Honkanen, P. Saarenrinne, T. Stoor, and J. Niinimaki. Recognition of highly overlapping ellipse-like bubble images. *Measurement Science and Technology*, 16(9):1760–1770, 2005.

[48] R. Horaud, B. Conio, O. Leboulleux, and B. Lacolle. An analytic solution for the perspective 4-point problem. *Computer Vision, Graphics, and Image Processing*, 47:33–44, 1989.

[49] Y. Hu, J. K. Liang, A. S. Myerson, and L. S. Taylor. Crystallization monitoring by Raman spectroscopy: Simultaneous measurement of desupersaturation profile and polymorphic form in flufenamic acid systems. *Industrial and Engineering Chemistry Research*, 44(5):1233–1240, 2005.

[50] E. J. Hukkanen and R. D. Braatz. Measurement of particle size distribution in suspension polymerization using in situ laser backscattering. *Sensors and Actuators B*, 96:451–459, 2003.

[51] D. P. Huttenlocher. Recognition by alignment. In A. K. Jain and P. J. Flynn, editors, *Three-dimensional object recognition systems*, Advances in image communication, pages 311–326. Elsevier, Amsterdam, 1993.

[52] D. P. Huttenlocher and S. Ullman. Recognizing solid objects by alignment with an image. *International Journal of Computer Vision*, 5(2):195–212, 1990.

[53] J.-H. Jang and K.-S. Hong. Fast line segment grouping method for finding globally more favorable line segments. *Pattern Recognition*, 35:2235–2247, 2002.

[54] P. Kahn, L. Kitchen, and E. Riseman. A fast line finder for vision-guided robot navigation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(11):1098–1102, November 1990.

[55] E. N. Kaufman and T. C. Scott. In situ visualization of coal particle distribution in a liquid fluidized bed using fluorescence microscopy. *Powder Technology*, 78:239–246, 1994.

[56] A. Kellerer. On the number of clumps resulting from the overlap of randomly placed figures in a plane. *Journal of Applied Probability*, 20(1):126–135, 1983.

[57] C. Lantuejoul. Computation of the histograms of the number of edges and neighbours of cells in a tessellation. In R. Miles and J. Serra, editors, *Geometrical Probability and Biological Structures: Buffon's 200th Anniversary*, number 23 in Lecture Notes in Biomathematics, pages 323–329, Berlin-Heidelberg-New York, 1978. Springer-Verlag.

[58] P. A. Larsen and J. B. Rawlings. Derivations for maximum likelihood estimation of particle size distribution using in situ video imaging. Technical Report 2007-01, TWMCC, Available at http://www.che.utexas.edu/twmcc/, March 2007.

[59] P. A. Larsen, J. B. Rawlings, and N. J. Ferrier. An algorithm for analyzing noisy, in situ images of high-aspect-aspect ratio crystals to monitor particle size distribution. *Chemical Engineering Science*, 61(16):5236–5248, 2006.

[60] P. A. Larsen, J. B. Rawlings, and N. J. Ferrier. Model-based object recognition to measure crystal size and shape distributions from in situ video images. *Chemical Engineering Science*, 62:1430–1441, 2007.

[61] G. Laslett. The survival curve under monotone density constraints with application to two-dimensional line segment processes. *Biometrika*, 69(1):153–160, 1982.

[62] F. Lewiner, J. P. Klein, F. Puel, and G. Fevotte. On-line ATR FTIR measurement of supersaturation during solution crystallization processes. Calibration and applications on three solute/solvent systems. *Chemical Engineering Science*, 56(6):2069–2084, 2001.

[63] M. Li and D. Wilkinson. Determination of non-spherical particle size distribution from chord length measurements. Part 1: Theoretical analysis. *Chemical Engineering Science*, 60(12):3251–3265, 2005.

[64] V. Liotta and V. Sabesan. Monitoring and feedback control of supersaturation using ATR-FTIR to produce an active pharmaceutical ingredient of a desired crystal size. *Organic Process Research & Development*, 8(3):488–494, 2004.

[65] D. G. Lowe. Three-dimensional object recognition from single two-dimensional images. *Artificial Intelligence*, 31(3):355–395, 1987.

[66] D. G. Lowe. Fitting parameterized three-dimensional models to images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(5):441–450, 1991.

[67] Z. Ma and G. Guiochon. Application of orthogonal collocation on finite elements in the simulation of non-linear chromatography. *Computers & Chemical Engineering*, 15(6):415–426, 1991.

[68] C. Mack. The expected number of clumps when convex laminae are placed at random and with random orientation on a plane area. *Proceedings of the Cambridge Philosophical Society*, 50:581–585, 1954.

[69] C. Mack. On clumps formed when convex laminae or bodies are placed at random in two or three dimensions. *Proceedings of the Cambridge Philosophical Society*, 52:246–256, 1956.

[70] D. L. Marchisio, R. D. Vigil, and R. O. Fox. Quadrature method of moments for aggregation–breakage processes. *Journal of Colloid and Interface Science*, 258(2):322–334, 2003.

[71] F. J. Massey. The Kolmogorov-Smirnov test for goodness of fit. *Journal of the Americal Statistical Association*, 46(253):68–78, March 1951.

[72] R. Miles. *Stochastic Geometry*, chapter On the elimination of edge effects in planar sampling, pages 228–247. John Wiley & Sons, 1974.

[73] Milken Institute. Biopharmaceutical industry contributions to state and U.S. economies. Available at www.milkeninstitute.org, 2004.

[74] S. M. Miller and J. B. Rawlings. Model identification and control strategies for batch cooling crystallizers. *AIChE Journal*, 40(8):1312–1327, August 1994.

[75] C. A. Mitchell, L. Yu, and M. D. Ward. Selective nucleation and discovery of organic polymorphs through epitaxy with single crystal substrates. *Journal of the American Chemical Society*, 123(44):10830–10839, 2001.

[76] S. Motz, S. Mannal, and E.-D. Gilles. Integral approximation—an approach to reduced models for particulate processes. *Chemical Engineering Science*, 59(5):987–1000, 2004.

[77] M. Naito, O. Hayakawa, K. Nakahira, H. Mori, and J. Tsubaki. Effect of particle shape on the particle size distribution measured with commercial equipment. *Powder Technology*, 100:52–60, 1998.

[78] T. Norris, P. K. Aldridge, and S. S. Sekulic. Determination of end-points for polymorph conversions of crystalline organic compounds using on-line near-infrared spectroscopy. *The Analyst*, 122(6):549–552, 1997.

[79] L. E. O'Brien, P. Timmins, A. C. Williams, and P. York. Use of in situ FT-Raman spectroscopy to study the kinetics of the transformation of carbamazepine polymorphs. *Journal of Pharmaceutical and Biomedical Analysis*, 36(2):335–340, 2004.

[80] T. Ono, J. ter Horst, and P. Jansens. Quantitative measurement of the polymorphic transformation of L-glutamic acid using in-situ Raman spectroscopy. *Crystal Growth and Design*, 4(3):465–469, 2004.

[81] B. O'Sullivan, P. Barrett, G. Hsiao, A. Carr, and B. Glennon. In situ monitoring of polymorphic transitions. *Organic Process Research & Development*, 7(6):977–982, 2003.

[82] D. B. Patience. *Crystal Engineering Through Particle Size and Shape Monitoring, Modeling, and Control*. PhD thesis, University of Wisconsin–Madison, 2002.

[83] D. B. Patience, P. C. Dell'Orco, and J. B. Rawlings. Optimal operation of a seeded pharmaceutical crystallization with growth-dependent dispersion. *Organic Process Research & Development*, 8(4):609–615, 2004.

[84] D. B. Patience and J. B. Rawlings. Particle-shape monitoring and control in crystallization processes. *AIChE Journal*, 47(9):2125–2130, 2001.

[85] L. Petzold. Differential/algebraic equations are not ODE's. *SIAM Journal on Scientific and Statistical Computing*, 3(3):367–384, September 1982.

[86] M. Pons, H. Vivier, K. Belaroui, B. Bernard-Michel, F. Cordier, D. Oulhana, and J. Dodds. Particle morphology: from visualization to measurement. *Powder Technology*, 103:44–57, 1999.

[87] S. L. Price. The computational prediction of pharmaceutical crystal structures and polymorphism. *Advanced Drug Delivery Reviews*, 56(3):301–319, February 2004.

[88] F. Puel, P. Marchal, and J. Klein. Habit transient analysis in industrial crystallization using two dimensional crystal sizing technique. *Chemical Engineering Research and Design*, 75(A2):193–205, 1997.

[89] J. B. Rawlings, S. M. Miller, and W. R. Witkowski. Model identification and control of solution crystallization processes: A review. *Industrial and Engineering Chemistry Research*, 32(7):1275–1296, July 1993.

[90] J. B. Rawlings, C. W. Sink, and S. M. Miller. Control of crystallization processes. In A. S. Myerson, editor, *Handbook of industrial crystallization*, pages 201–230. Butterworth-Heinemann, USA, 2nd edition, 2002.

[91] S. Roach. *The Theory of Random Clumping*. Methuen's monographs on applied probability and statistics. Methuen & Company, London, 1968.

[92] S. Rohani and G. Zhang. On-line optimal control of a seeded batch cooling crystallizer. *Chemical Engineering Science*, 58(9):1887–1896, 2003.

[93] A. L. Rohl. Computer prediction of crystal morphology. *Current Opinion in Solid State & Material Science*, 7(1):21–26, 2003.

[94] A. Ruf, J. Worlitschek, and M. Mazzotti. Modeling and experimental analysis of PSD measurements through FBRM. *Particle and Particle Systems Characterization*, 17:167–179, 2000.

[95] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman. LabelMe: a database and web-based tool for image annotation. MIT AI Lab Memo, AIM-2005-025, September 2005.

[96] K. Sakamoto and R. W. Rousseau. Sizing elongated crystals using a width distribution function: Application to aspartame. *Industrial and Engineering Chemistry Research*, 39:3949–3952, 2000.

[97] J. Sanyal, D. L. Marchisio, R. O. Fox, and K. Dhanasekharan. On the comparison between population balance models for CFD simulation of bubble columns. *Industrial and Engineering Chemistry Research*, 44(14):5063–5072, 2005.

[98] Semiconductor Industry Association. Total semiconductor world market sales and shares. Available at www.sia-online.org, 2005.

[99] L. Shen, X. Song, M. Iguchi, and F. Yamamoto. A method for recognizing particles in overlapped particle images. *Pattern Recognition Letters*, 21(1):21–30, January 2000.

[100] D. Shi, N. H. El-Farra, M. Li, P. Mhaskar, and P. D. Christofides. Predictive control of particle size distribution in particulate processes. *Chemical Engineering Science*, 61(1):268–281, 2006.

[101] G. W. Snedecor and W. G. Cochran. *Statistical Methods*. Iowa State University Press, Ames, Iowa, 8 edition, 1989.

[102] H. Solomon. *Geometric Probability*. SIAM Publications, Philadelphia, PA, 1978.

[103] C. Starbuck, A. Spartalis, L. Wai, J. Wang, P. Fernandez, C. M. Lindemann, G. X. Zhou, and Z. Ge. Process optimization of a complex pharmaceutical polymorphic system via in situ Raman spectroscopy. *Crystal Growth and Design*, 2(6):515–522, 2002.

[104] I. Svensson, S. Sjostedt-De Luna, and L. Bondesson. Estimation of wood fibre length distributions from censored data through an EM algorithm. *Scandinavian Journal of Statistics*, 33:503–522, 2006.

[105] T. Togkalidou, M. Fujiwara, S. Patel, and R. D. Braatz. Solute concentration prediction using chemometrics and ATR-FTIR spectroscopy. *Journal of Crystal Growth*, 231(4):534–543, 2001.

[106] M. J. Van Der Laan. The two-interval line-segment problem. *Scandinavian Journal of Statistics*, 25:163–186, 1998.

[107] E. W. Van Zwet. Laslett's line segment problem. *Bernoulli*, 10(3):377–396, 2004.

[108] J. Villadsen and M. L. Michelsen. *Solution of Differential Equation Models by Polynomial Approximation*. Prentice-Hall, Englewood Cliffs New Jersey, 1978.

[109] U. Vollmer and J. Raisch. Population balance modelling and H-infinity—controller design for a crystallization process. *Chemical Engineering Science*, 57(20):4401–4414, 2002.

[110] F. Wang, J. A. Wachter, F. J. Antosz, and K. A. Berglund. An investigation of solvent-mediated polymorphic transformation of progesterone using in situ Raman spectroscopy. *Organic Process Research & Development*, 4(5):391–395, 2000.

[111] I. Weissbuch, L. Addadi, M. Lahav, and L. Leiserowitz. Molecular recognition at crystal interfaces. *Science*, 253(5020):637–645, 1991.

[112] I. Weissbuch, M. Lahav, and L. Leiserowitz. Toward stereochemical control, monitoring, and understanding of crystal nucleation. *Crystal Growth and Design*, 3(2):125–150, 2003.

[113] B. Wijers. *Nonparametric estimation for a windowed line-segment process*. Stichting Mathematisch Centrum, Amsterdam, Netherlands, 1997.

[114] D. Winn and M. Doherty. A new technique for predicting the shape of solution-grown organic crystals. *AIChE Journal*, 44(11):2501–2514, 1998.

[115] K. C. Wong and J. Kittler. Recognition of polyhedral objects using triplets of projected spatial edges based on a single perspective image. *Pattern Recognition*, 34:561–586, 2001.

[116] J. Worlitschek, T. Hocker, and M. Mazzotti. Restoration of PSD from chord length distribution data using the method of projections onto convex sets. *Particle and Particle Systems Characterization*, 22(2):81–98, August 2005.

[117] J. Worlitschek and M. Mazzotti. Model-based optimization of particle size distribution in batch-cooling crystallization of paracetamol. *Crystal Growth and Design*, 4(5):891–903, 2004.

[118] M. Wulkow, A. Gerstlauer, and U. Nieken. Modeling and simulation of crystallization processes using PARSIVAL. *Chemical Engineering Science*, 56(7):2575–2588, 2001.

[119] R. Xu and O. A. D. Guida. Comparison of sizing small particles using different technologies. *Powder Technology*, 132:145–153, 2003.

[120] L. X. Yu, R. A. Lionberger, A. S. Raw, R. D'Costa, H. Wu, and A. S. Hussain. Applications of process analytical technology to crystallization processes. *Advanced Drug Delivery Reviews*, 56(3):349–369, 2004.

[121] Q. Yu and N.-H. L. Wang. Computer simulations of the dynamics of multicomponent ion exchange and adsorption in fixed beds – gradient-directed moving finite element method. *Computers & Chemical Engineering*, 13(8):915–926, 1989.

Vita

Paul A. Larsen was born in Blackfoot, Idaho to Stephen and Susan Larsen. In June 1996, he graduated as valedictorian of his class from Snake River High School in Blackfoot. It has all been downhill since then.

Permanent Address: 4909 Ascot Lane #R1
Madison, WI 53711

This dissertation was prepared with LaTeX 2$_\varepsilon$[1] by the author.

[1]This particular University of Wisconsin complient style was carved from The University of Texas at Austin styles as written by Dinesh Das (LaTeX 2$_\varepsilon$), Khe–Sing The (LaTeX), and John Eaton (LaTeX). Knives and chisels wielded by John Campbell and Rock Matthews.