Direct Image Feature Extraction and Multivariate Analysis for Crystallization Process Characterization

Frederik J. S. Doerr, Cameron J. Brown, and Alastair J. Florence*

ABSTRACT: Small-scale crystallization experiments (1–8 mL) are widely used during early-stage crystallization process development to obtain initial information on solubility, metastable zone width, as well as attainable nucleation and/or growth kinetics in a material-efficient manner. Digital imaging is used to monitor these experiments either providing qualitative information or for object detection coupled with size and shape characterization. In this study, a novel approach for the routine characterization of image data from such crystallization experiments is presented employing methodologies for direct image feature extraction. A total of 80 image features were extracted based on simple image statistics, histogram parametrization, and a series of targeted image transformations to assess local grayscale characteristics. These features were utilized for applications of clear/cloud point detection and crystal suspension density prediction. Compared to commonly used transmission-based methods (mean absolute error 8.99 mg/mL), the image-based detection method is significantly more accurate for clear and cloud point detection with a mean absolute error of 0.42 mg/mL against a manually assessed ground truth. Extracted image features were further used as part of a partial least-squares regression (PLSR) model to successfully predict crystal suspension densities up to 40 mg/mL ($R^2 > 0.81$, $Q^2 > 0.83$). These quantitative measurements reliably provide crucial information on composition and kinetics for early parameter estimation and process modeling. The image analysis methodologies have a great potential to be translated to other imaging techniques for process monitoring of key physical parameters to accelerate the development and control of particle/crystallization processes.

1. INTRODUCTION

Crystallization is a purification and separation process widely applied in the agricultural, pharmaceutical, and chemical industries during the production of specialty or high-value chemicals. The successful development, optimization, and control of crystallization processes rely on the accurate determination of solid—liquid phase equilibria, system thermodynamics, rate process kinetics, and solid-state chemistry. Process design decisions can have a direct impact on the robustness and stability of the crystallization process, production yields, purity, as well as downstream isolation and performance attributes of the crystalline material, which could influence the overall therapeutic efficacy as part of the formulated final product.

Process analytical technology (PAT) is increasingly applied for the analysis, monitoring, and control of pharmaceutical manufacturing processes with guidance from regulatory agencies encouraging the integration of PAT during process implementation. In the context of crystallization, strategies for the quantification of crystal suspension densities are mostly derived from solute concentration measurements using spectroscopic techniques such as UV/vis, Raman, or infrared. Focused beam reflectance measurements (FBRM) can be used to monitor particle size distribution (PSD) trajectories in situ avoiding the need for sampling, sample transfer, and offline measurements. However, the chord length distribution data from FBRM needs to be transformed to access the PSD and is therefore often only used semi-quantitatively to assess relative changes in the particle counts across highly discretized size fractions. More recently, techniques including optical imaging and FBRM have been used to successfully quantify particle size and suspension densities up to 10 wt %; however, this approach combines measurements from multiple PAT probes, which is not feasible for small-scale experiments.

Optical imaging itself is a popular PAT application to monitor manufacturing processes involving solid—liquid, mixed phase systems such as particle suspensions during crystallization. Combined with automated image processing and analysis methodologies, data from optical imaging can be exploited to extract quantitative information on relevant particle properties.
related to size and shape. Various commercial PAT systems are available for process imaging with probes directly submerged into the process stream to enable imaging of suspended particles. These off-the-shelf commercial systems typically include proprietary image processing and analysis software for size and shape quantification. Image analysis methods to detect individual crystals/particles during image processing and to extract size and shape information on suspended particles are commonly limited in their applicability to low suspension densities where the edges of individual particles can still be easily detected, and the effects of extensive particle–particle overlapping in the images remain negligible. Object detection also commonly relies on image processing methods for noise reduction, e.g., local smoothing using image filters or despeckling after binarization, which can significantly impact the quantification accuracy of sensitive particle descriptors and which reduce the accuracy of these image analysis methods to detect early nucleation where only a few small crystalline particles might be present inside the image focal plane. There are examples using simple image statistics as a basis for process monitoring and quantification without the need for extensive image processing. In these cases, local concentrations or suspension densities are directly correlated to individual, extracted image pixel intensity statistics, most commonly the mean gray level intensity. This approach however has not yet been further expanded to include a wider range of optimized image features which might improve its ability to resolve physical changes in the system.

Small-scale experiments (1–8 mL) are often used during early-stage process development to obtain basic information on solute solubility, metastable zone width, as well as estimates on nucleation and growth kinetics. The availability of commercial, automated, and parallel reactor systems allows such experiments to be carried out at the milliliter scale in a material sparing manner. However, the small dimensions of the vessels limit the use of high-end PAT equipment designed for applications in larger-scale crystallizers. Instead, transmission measurements are commonly used to monitor each experiment and detect clear and cloud points related to full solid dissolution or initial nucleation/precipitation, respectively.

Given the value of accurate data describing the thermodynamics and kinetics of crystal formation at small scale, an analysis framework based on image features has been developed which can be employed to complement traditional object detection methods and enhance the information content accessible through standard optical imaging techniques. A total of 80 image features are part of this data-driven computer vision approach and were extracted from each image using targeted image transformations, parametrization of the pixel intensity histogram, and basic image intensity statistics. These methodologies were utilized to analyze a series of image data sets from small-scale experiments of mefenamic acid in solvent mixtures of diglyme and water (70:30–90:10, w/w). The image analysis framework was applied (A1) to support the accurate detection of early nucleation (cloud point) and complete dissolution at the solubility line (clear point) as well as (A2) to provide estimates of crystal suspension densities using extracted image features as input variables for a partial least-squares regression (PLSR) model. The work further includes an experimental design proposal for in-sample PLSR calibration using a stepwise heating protocol.

2. MATERIALS AND EXPERIMENTAL METHODS

2.1. Materials. Mefenamic acid (MFA) was sourced from Sigma (Merck KGaA, Darmstadt, Germany, Lot#MKCH3607). Suspensions with changing MFA weight ratios were prepared in a 5 g mixture of diglyme (DIG, bis(2-methoxyethyl) ether) and deionized water (WAT) with ratios of 70:30, w/w (DIG70), 80:20, w/w (DIG80) and 90:10, w/w (DIG90). DIG was sourced from Fisher Scientific (ACROS Organics, Waltham, United States). WAT was purified and deionized using a Milli-Q system (Merck KGaA, Germany). Details of all prepared samples for this study are tabulated in Table 1.

| experiment | DIG [wt %] | WAT [wt %] | $m_{\text{DIG}}$ [mg] | $m_{\text{WAT}}$ [mg] | $x_{\text{MFA}}$ [g/L] |
|------------|-----------|------------|----------------------|----------------------|----------------------|
| MFA-DT     | 70        | 30         | 262.4                | 5                    | 0.052                |
| MFA-70-1   | 70        | 30         | 262.4                | 5                    | 0.052                |
| MFA-70-2   | 70        | 30         | 190.5                | 5                    | 0.052                |
| MFA-70-3   | 70        | 30         | 133.7                | 5                    | 0.052                |
| MFA-70-4   | 70        | 30         | 70.3                 | 5                    | 0.052                |
| MFA-80-1   | 80        | 20         | 391.3                | 5                    | 0.052                |
| MFA-80-2   | 80        | 20         | 301.3                | 5                    | 0.052                |
| MFA-80-3   | 80        | 20         | 221.0                | 5                    | 0.052                |
| MFA-90-1   | 90        | 10         | 748.5                | 5                    | 0.052                |
| MFA-90-2   | 90        | 10         | 601.5                | 5                    | 0.052                |
| MFA-90-3   | 90        | 10         | 502.9                | 5                    | 0.052                |
| MFA-90-4   | 90        | 10         | 421.2                | 5                    | 0.052                |
| MFA-90-5   | 90        | 10         | 363.5                | 5                    | 0.052                |

2.2. Crystalline. Small-scale crystallization experiments were conducted in 8 mL clear glass vials (height 61 mm × diameter 16.6 mm, hydrolytic class 1) inserted in a Crystalline (Technobis Crystallization Systems, Alkmaar, The Netherlands). Mixing was provided through magnetic stirring with a bottom stirring speed of 700 rpm using a PTFE-coated elliptical stirrer (dimensions 10 × 3 mm). In general, fast and homogeneous mixing of most crystal suspensions can be assumed for these small-scale experiments. The Crystalline platform itself was equipped with a temperature controller, a transmissivity sensor, and an imaging system (RR-PV module). Reactor jacket temperature and transmission measurements were recorded at 1 Hz. A tuning step was applied to the transmission signal when the solution reaches clear point (i.e., crystals are fully dissolved). This essentially sets the laser power to 100% transmission. The transmission signal was further processed to reduce noise using a Hampel filter with a moving window size ($k_{\text{Hamp}}$) of 7 and an outlier criterion of $n_{\text{out}} = 3$ local standard deviations as well as a moving average filter (Savitsky-Golay, order 1) with a moving window size ($k_{\text{SG}}$) of 9. For the processed transmission signal, clear and cloud points were detected at a threshold of >99.9% transmissivity$_{\text{sm}}$ and <99.9% transmissivity$_{\text{mur}}$ respectively. The imaging frequency was user-defined with frame-rates between 0.5 and 0.05 fps depending on the expected process dynamics during crystallization. At the end of each experiment, all process data and images with a resolution of 480 × 640 px at an image pixel size of 2.8 μm/px were exported for offline data processing.

Temperature profiles for the small-scale experiments were designed for image feature calibration and to characterize basic crystallization (thermo)dynamics during cooling and heating. An example of the implemented Crystalline temperature profile for each vessel is shown in Figure 1 with additional details for each experiment provided in the Supporting Information (Table S1). Each of the three phases of the applied profile, P1–3 is described further. (P1) Each experiment was designed to undergo an initial stepwise heating sequence with heating steps of 5 K and isothermal data collection at holding periods of 30 min. Images for PLSR calibration were collected at the end of each isothermal temperature stage in P1 after an equilibrium time of 5 min. The calibrated PLSR model was used to predict crystal suspension
densities during dynamic crystallization events. A constant cooling/heating rate of ±0.5 K was used for (P2) metastable zone width assessment and (P3) solubility determination. For MFA-DT, isothermal holding periods were extended to 120 min to create an image data set for initial method development.

2.3. Image Processing and Analysis. All routines for automated image processing and analysis were implemented in MATLAB.

Figure 1. Crystalline experiments with programmed temperature profile. (P1) Step-wise heating sequence (heating 1 K/min) with 30 min isothermal conditions for in-sample PLSR calibration. (P2) Metastable zone width experiment with crystal nucleation detection (slow cooling ~0.5 K/min) and (P3) dissolution profile for solubility temperature detection (slow heating at 0.5 K/min). bk (annotated with yellow arrow) shows where background image data were collected at the end of P1 and P3.

Figure 2. Time-resolved (green) transmission signal compared to selected (magenta) image feature signals during stepwise heating (red line, temperature): (A) mean image intensity (IntM), (B) HELM$_5$, and (C) WAVR. Manually identified clear point (blue dashed, complete dissolution).
Multicore, parallel processing of individual 2D images was used to accelerate routine image analysis. Image feature extraction focuses on the quantification of image characteristics from eight-bit grayscale image data collected during each experimental run. A total of 80 image features are extracted using three distinct approaches: (1) directly from the image raw data based on global image intensity statistics \( (n = 6) \), (2) through a parametrization of the pixel intensity histogram \( (n = 15) \), and (3) from a variety of targeted image transformations or local image variance analyses \( (n = 59) \). Image transformations provide an excellent opportunity to explore a diversity of image features for target applications, e.g., by adjusting kernel parameters which were optimized in the context of crystal suspension characterization. An overview of all 80 extracted image features with additional details on the employed parameters is provided in Table S2 (Supporting Information).

**2.4. Partial Least Square Regression (PLSR)**. Partial least square regression (PLSR) analysis was employed for crystal suspension density prediction using the extracted image features. The crystal suspension density \( (x_c) \) is defined here as the weight of suspended crystalline solids per volume element of the solvent liquid phase due to the nature of the data acquisition approach imaging a fixed suspension volume but with changing solvent compositions (DIG70–DIG90). Effects of solute molecules in the liquid phase and the contribution of crystalline material to the overall suspension density are assumed to be equal across all solvent systems.

An implementation of the SIMPLS algorithm was used for PLSR which calculates the PLSR factors directly as linear combinations of the original variables and therefore retains good interpretability for future efforts on feature selection and/or expansion. For PLSR, image feature data from isothermal conditions during an initial stepwise heating procedure were isolated (Figure 1 P1). Discrete binning and averaging were used to assess prediction improvements by reducing image-to-image variability. A bin size of \( k_{av} = 5 \) images was selected to compensate for potential image-to-image variability while retaining good time resolution during dynamic crystallization events. The image feature data were mean-centered prior to PLSR model training and testing. The number of latent variables \( (n_{comp}) \) was optimized based on the calculated mean square error of the prediction during cross-validation (MSECV). The PLSR model performance to accurately predict crystal suspension densities was assessed against (I) potential image-to-image variability and (II) data set-to-data set variability. (I) Feature data from individual images of a single image data set, MFA-DT, were randomly divided into training and test data (85:15). The training data were used...
to optimize the number of latent variables (ncomp) during 10-fold cross-validation. (II) Individual image data sets (MFA-70-1−MFA-90-5) were randomly divided into training and test data sets (10:2). The training data sets were used to optimize the number of latent variables (ncomp) during four-fold cross-validation.

3. RESULTS AND DISCUSSION

3.1. Image Feature Extraction and Statistics. Image features were extracted using 31 methods aiming to quantify image properties from basic image intensity statistics (n = 6), from a parametrization of the pixel intensity histogram (n = 15) and from image transformations operating within a local gray level pixel neighborhood (n = 59). Features extracted from targeted image transformations, e.g., using Prewitt23 or Sobel24 operators were of particular interest. These image transformations are often employed for edge detection in computer vision applications and therefore might be highly sensitive to physical changes in the crystal suspension. Many of the methods tested in this study provide multiple image features or metrics which might include mean pixel intensity, variance, or entropy from each of the transformed images. In total, 80 image features

Figure 4. Pair-wise assessment of peak resolution for all image feature data distributions to distinguish between discrete suspension densities (x_c) between 37.8 and 0.0 mg/mL (S = 0.7). (A) Three examples of WAVR comparing data distribution metrics and peak resolution. Dashed line indicates measured signal distribution at μ ± 2σ. Peak resolution of all features are shown in (B) for feature data extracted from raw images and (C) after using a discrete binning and image averaging (k_av = 5) for noise reduction. Black squares indicate the ratio of image features with a peak resolution higher than 1 (n_R > 1/n).
were extracted and assessed. Three selected feature signals are visualized in Figure 2 to illustrate the complexity of the data analysis problem but also opportunities for correlating image features with physical changes in the sample during a stepwise heating profile. These include (A) the mean image intensity of the raw data (IntM) often used during image analysis as a basic image feature, (B) HELM, assessing pixel intensity variance against the local mean background used for clear and cloud point detection, and (C) WAVR from a 2-D wavelet decomposition with high peak resolution across changing suspension densities. Each feature signal exhibits unique characteristics related to their temporal signal intensity and variance.

Most image transformations employ application-specific kernels which operate within a defined pixel neighborhood. The kernel shape, size, and values can be optimized to capture image attributes of interest, e.g., for texture or pattern recognition. For crystal suspensions, for example, Figure 3 visualizes the impact of changing kernel sizes for a range filter with a kernel size of (Figure 3B, RngS) $7 \times 7$ px and (Figure 3C, RngS1) $71 \times 71$ px, respectively. While a smaller kernel size of $7 \times 7$ px can be used for edge enhancement to get an indication of the number of crystalline objects, larger kernel sizes of $71 \times 71$ px are very sensitive to local intensity fluctuations in the presence of only a few suspended crystalline objects. Therefore, this filter seems particularly useful for the characterization of images from crystallization experiments with medium and dilute crystal suspension densities and for early detection of particle formation, respectively. Other image filters follow similar methodologies and can be tailored using user-defined parameters which were partially preoptimized based on collected crystallization images in the MFA-DT data set. An overview of all methods for image feature extraction and additional details on the image filter kernel parameters are provided in Table S2 (Supporting Information).

The performance of each extracted image feature to resolve physical changes during crystallization was assessed and compared through a calculation of the peak resolution (Res) between pairs of collected measurement distributions with known, discrete changes in the crystal suspension density. Res was calculated as shown in eq 1 assuming normal distribution:

$$\text{Res}_n = \frac{|\mu_n - \mu_{n+1}|}{\sqrt{4\sigma_n^2 + 4\sigma_{n+1}^2}}$$

where $\mu$ and $\sigma$ are the mean and the standard deviation of the image feature distributions, respectively, for samples $n$ and $n+1$. When Res is equal to 1, 95% of the measured feature values can be correctly classified between the two samples. Therefore, it is desirable to maintain a Res value greater than 1 across all changes in crystal suspension density.

Image feature resolution was assessed using the MFA-DT image data set with 11 distinct isothermal temperature levels between 30°C ($x_c = 37.8$ mg/mL) and 80°C ($x_c = 0.0$ mg/mL, $S = 0.7$) during a stepwise heating protocol (example shown in Figure 1 P1). Expected crystal suspension densities were calculated based on the MFA phase diagram presented in Figure 6. During each isothermal period, an average of 239 images were collected. Figure 4A shows example measurement distributions with pairwise resolution assessment for WAVR. The results indicate that between each crystal suspension density pair (magenta versus cyan) the distributions of WAVR values are sufficiently different to distinguish between them, which is reflected in a Res value greater than 1 for all cases. Figure 4B expands upon this concept and shows the resolution values across all features for the raw image data and in Figure 4C the resolution values after using discretized binning to reduce image feature noise ($k_{av} = 5$). The black trend line indicates the ratio of image features that achieve a Res > 1.

In general, the ability to distinguish changes in the crystal suspension density based on image features in MFA-DT is not uniform but strongly depends on the crystal suspension density itself. Interestingly, images with very high suspension density ($x_c > 23.1$ mg/mL) show the highest resolution between the collected measurement distributions and maintain a Res > 1.

Figure 4B. At suspension densities between 23.1 and 4.3 mg/mL, the peak resolution indicates that the distribution of image feature values is insufficiently separated to distinguish between discrete changes in the crystal suspension density. This is likely due to having few particles in suspension, resulting in a reduced probability of particles appearing in these images and an increased influence related to the size and shape of each individual particle itself. Feature resolution in this range with more dilute suspensions can be improved using discretized binning, Figure 4C. Differences between $x_c = 4.3$ mg/mL and $x_c = 0.0$ mg/mL ($S = 0.9$, no particles), i.e., compare image feature signals above and below the MFA solubility, show a good peak resolution at this crucial point during dissolution (clear point) with 65% of all image features exhibiting a resolution of Res > 1.

A ranking of all image features with numeric values from the resolution assessment is provided in Table S3 and Table S4 (Supporting Information) for the raw data and after discretized binning for noise reduction ($k_{av} = 5$), respectively. Each table further includes the determined Pearson correlation coefficient (PCC) to quantify the overall linear correlation between each image feature and the calculated crystal suspension density as the response variable. Among them, WAVR (rank 1) exhibits a high positive linear correlation (PCC = 0.99) with a consistent peak resolution of Res > 1 across the full range of suspension densities between $x_c = 37.8$ mg/mL and $x_c = 0.0$ mg/mL ($S = 0.9$). Notably, a number of image features outperform the basic mean image intensity (IntM, rank 19, Res($k_{av} = 5$) = 2.16 ± 1.10 $\in [0.61, 4.28]$, PCC = −0.94) and the transmission signal (not ranked, Res($k_{av} = 5$) = 0.93 ± 1.09 $\in [0.01, 2.86]$, PCC = −0.69) in terms of their ability to resolve physical changes in the crystal suspension density especially for high suspension densities up to 37.80 mg/mL, confirming the value of advanced image parameterization and targeted feature engineering efforts. Example images for each supersaturation level under isothermal conditions are shown in Figure S1 and Figure S2 (Supporting Information), respectively. The histogram for the images in Figure S2 was adjusted (stretched) to visualize subtle differences in the image intensities for the reader, which are quantified during image feature extraction but can otherwise not be directly observed in the raw images.

The performance of individual image features might be impacted through the optical properties of the crystalline material in this study. Additionally, differences in the particle size distribution and the particle shape could affect image feature performance where size and shape features are significantly larger than the image pixel size. Therefore, the performance of individual image features as part of this data-driven approach for crystallization image analysis needs to be re-evaluated for new compound systems. Depending on the desired application, features with low resolution and low correlation to real physical changes in the system can be excluded to accelerate image data...
analysis for a routine or real-time implementation. Compared to traditional object detection methods to determine number density, size, and shape information, extracted image features are not inherently meaningful. The following two applications aim to further explore the use of extracted image features for (A1) detection of complete dissolution (clear point) or crystal nucleation (cloud point) and (A2) crystal suspension density prediction.

3.2. Application (A1) - clear and cloud Point Detection.

Accurate clear and cloud point detection is crucial to obtain information on the compound solubility and metastable zone width which are often used as the basis for process development and optimization. Small-scale crystallization experiments typically rely on transmission measurements to automatically detect complete dissolution (clear point) or initial crystal nucleation (cloud point) using a user-defined, fixed threshold level. This first application explores the use of extracted image features as an alternative method to transmission to further improve clear and cloud point detection accuracy.

A rational method for feature selection is essential to identify robust image feature(s) for clear and cloud point detection. A simple single feature threshold might be used as a detection criterion to facilitate rapid, routine analysis. The median background value of each image feature \( m_y \) was calculated using clear background images of the solution. Consequently, cloud points were detected once the feature signal diverges from \( m_y \pm n \sigma_y \) for clear point detection, the inverse objective, i.e. a convergence criterion, was used instead.

Feature signal processing can help to increase the signal-to-noise ratio (SNR) of extracted image features and, therefore, improve the overall accuracy and precision of detection. Two filters were selected for preprocessing the extracted raw signals of all 80 image features: (1) a Hampel filter for local outlier detection caused by individual images with strong image feature fluctuations, e.g., related to foreign particles or insoluble impurities in the suspension and (2) a moving average filter for local smoothing of random noise (Savitzky-Golay, order 1). Three filter parameters, the optimum window size of each moving filter \( k_{Hpl} \) and \( k_{SG} \), and the outlier criterion for the Hampel filter \( n_{Hpl,\sigma} \) (number of local standard deviations) were subsequently optimized against a user-defined ground truth for

![Figure 5. Comparison of (A) clear point detection and (B) cloud point detection](https://doi.org/10.1021/acs.cgd.1c01118)
three randomly selected image data sets (MFA-70-3, MFA-80-2, and MFA-90-4) with the objective to minimize the temperature mean squared error (MSE) for clear and cloud point detection. For the detection criterion, the divergence/convergence threshold was defined at three different levels with $n_{\text{nuc}}$ of 4, 8, and 16 to further evaluate the overall robustness of this approach.

HELM$_5$ was identified as the best performing image feature for clear and cloud point detection with an overall MSE of 0.05 °C$^2$ after preprocessing the image feature data with $k_{\text{bpl}} = 29$, $n_{\text{HPL,}0} = 0.1$, and $k_{\text{SG}} = 11$. Details about the performance of all features for clear and cloud point detection and feature-specific optimized signal preprocessing parameters ($k_{\text{bpl}, \text{HPL,}0}$ and $k_{\text{SG}}$) are provided in Table S5 (Supporting Information). Among them, BREN (Brenner’s focus) and PRWV (Previtt operator) also show excellent performance. In contrast to HELM$_5$, which aims to quantify absolute differences related to its background intensity, BREN and PRWV are based on an assessment of the local spatial intensity gradient where they are traditionally used to assess image focus or for edge detection, respectively. Low performing image features are related to global image intensity statistics such as parameters from the image intensity histogram (HistSpan, HistSumBK) or the image intensity index of dispersion (IntIdxD). These image features do not operate within defined image pixel neighborhoods and, therefore, are less likely to distinguish random global image intensity fluctuations from highly localized intensity fluctuations due to individual crystals in the imaging field of view. Figure 5 shows the application of HELM$_5$ for MFA-80-2 during (A) slow heating at 0.5 K/min and (B) slow cooling at −0.5 K/min in direct comparison to a transmission-based clear and cloud point detection method. The processed, smoothed feature signal (black dashed line) shows significantly reduced local fluctuations compared to the feature raw signal (magenta line); however, the strong performance of HELM$_5$ for clear and cloud point detection is related to its consistent background signal of $1.00 \pm 6.71 \times 10^{-5}$ ($S = 0.7$–0.9). The consistent background signal of HELM$_5$ is related to the nature of this image feature which quantifies the average ratio between pixel intensities and the background mean gray level intensity of its pixel neighborhood (for HELM$_5$, a $5 \times 5$ pixel kernel), therefore, compensating the impact of local random noise and small intensity fluctuations. Image series in Figure 5C,D are provided to give an impression of the dissolution and nucleation dynamics. For the transmission signal (green box), the detected clear point with a threshold of >99.9% and cloud point with a threshold of <99.9% (blue dotted pictures frames) after signal processing (transmissivity$_{\text{raw}}$) are shown alongside images from intermediate time points to visualize the progress of crystal dissolution and crystal nucleation/growth, respectively. The image-based detection method shows an excellent performance with a detected clear point at 51.7 °C (complete dissolution, solubility) and a cloud point at 36.0 °C (initial nucleation, metastable zone width) in stark contrast to the transmission signal with a significantly reduced sensitivity to detect low crystal suspensions densities providing highly misleading clear/cloud point values.

Figure 6 further shows the impact of both clear and cloud point detection methods on the phase diagram of MFA constructed from a series of 12 small-scale experiments (MFA-70-1–MFA-90-5). Image feature-based clear and cloud points (Figure 6 filled symbols) are consistently shifted to higher temperatures, indicating later detection of complete dissolution during heating and earlier initial nucleation detection during cooling compared to a detection based on the transmission signal with a 99.9% threshold (Figure 6 empty symbols). Across all data sets (MFA-70-1–MFA-90-5), the average difference between both methods, HELM$_5$ image feature versus transmission signal, is $6.63 \pm 3.61 ^\circ \text{C} \in [3.10 ^\circ \text{C}, 14.90 ^\circ \text{C}]$ and $2.59 \pm 1.56 ^\circ \text{C} \in [0.00 ^\circ \text{C}, 5.10 ^\circ \text{C}]$ for clear and cloud point detection, respectively. Image feature-based clear and cloud points are
consistent against a manually assessed, user-defined ground truth (filled markers with black edge).

The van’t Hoff equation for nonideal solutions (eq 2) was used to interpret the experimental solubility data (clear points) where $x$ is the mole fraction of the solute in the solution, $T$ is the temperature, $R$ is the ideal gas constant, $\Delta H_d$ is the enthalpy of dissolution, and $\Delta S_d$ is the entropy of dissolution.\(^1\)

\[
\ln(x) = -\frac{\Delta H_d}{RT} + \frac{\Delta S_d}{R}
\]

Point estimates for $\Delta H_d$ and $\Delta S_d$ from a regression analysis of each individual solvent composition were used to generate solubility curves in Figure 6 for (dashed line) transmission-based clear points, (dash-dotted line) image feature-based clear points and (solid line) the user-defined ground truth. Additional details of the regression analysis are provided in Figure S3 (Supporting Information).

In the context of the van’t Hoff solubility model, one can distinguish and quantify the contribution of systematic and random measurement errors related to the overall accuracy and precision of the detection method. The results indicate a good fit of eq 2 to the experimental data with comparatively low unexplained, random error contributions for each of the applied detection methods ($R^2 \geq 0.99$, residual mean squared error (RMSE) $\leq 3.82 \times 10^{-5}$). The systematic error (bias) of each detection method was assessed using the estimated solubility model parameters $\Delta H_d$ and $\Delta S_d$ (user-defined ground truth, $\Delta H_d \in [20.05 \text{ kJ/mol, 28.14 kJ/mol}], \Delta S_d \in [34.09 \text{ J/(mol K), 42.92 J/(mol K)]}$), which are both increasing with higher solvent ratios of water, consistent with other studies investigating MFA solubility in water and a range of organic solvents.\(^3\) Across all solvent compositions, the method for image feature-based solubility detection has an error of $0.00 \pm 0.68 \text{ kJ/mol}$ and $-0.20 \pm 2.09 \text{ J/(mol K)}$ for $\Delta H_d$ and $\Delta S_d$ respectively. In contrast, the transmission-based solubility detection method has a significantly higher systematic error of $-2.90 \pm 3.59 \text{ kJ/mol}$ and $-7.38 \pm 10.38 \text{ J/(mol K)}$ for $\Delta H_d$ and $\Delta S_d$ respectively.

On the basis of the experiments performed in this study ($y_{\text{c}}, n = 12$) and using the predicted solubility of the established MFA phase diagram ($\hat{y}_{\text{Truth}}(T)$), the observed mean absolute detection error can be expressed as the mean crystal suspension density at the detected clear point for each detection method ($\text{MAE} = y_i - \hat{y}_{\text{Truth}}$ Figure 6 red lines). The image feature-based approach has an MAE of $<0.001 \pm 0.001 \text{ g/g} \in [-0.001, 0.002]$ or $0.02 \pm 1.16 \text{ mg/mL} \in [-1.13, 2.00]$ and, therefore, aligns well with the manually determined clear points (ground truth).

For the transmission-based clear points detected at a 99.9% transmission threshold, the MAE is $0.009 \pm 0.003 \text{ g/g} \in [0.006, 0.016]$ or $8.99 \pm 2.92 \text{ mg/mL} \in [5.86, 15.26]$, indicating an incorrect, early detection during constant heating at 0.5 K/min with an extensive crystal mass still present in the sample. This confirms initial qualitative observations presented in Figure 5C (green box) showing significant amounts of crystalline particles at the detected clear point using the transmission-based method. Similar observations with an improved determination of the clear point using image-based detection have been described in the literature after collected image data sets from small-scale crystallization experiments were manually reviewed and compared.\(^38,39\)

Incorrect clear and cloud point detection significantly affects the determined compound solubility or metastable zone width and therefore crystallization process design decisions through error propagation further emphasizing the importance of accurate and precise metrics for clear and cloud point detection. The results have demonstrated that image features such as
HELMs can be used as an alternative to transmission-based detection methods and provide a more reliable determination of these basic (thermo)dynamic properties for new compounds during early-stage, small-scale crystallization process development.

3.3. Application (A2) - Suspension Density Prediction. Crystal suspension density measurements are useful for crystallization model development to inform system dynamics during parameter estimation or can be used for direct crystallization process monitoring and control. However, a quantification of the crystal suspension density is not readily accessible via direct measurement during small-scale, highly parallelized experiments. This second application aims to explore the use of image features for the prediction of crystal suspension densities from crystallization image data. Crystal suspension densities are defined here as the weight of suspended crystalline solids per volume of the solvent liquid phase ($\chi_v$ in mg/mL) based on the nature of the data acquisition approach imaging a fixed suspension volume but with changing solvent compositions (DIG70–DIG90).

PLSR analysis is a multivariate analysis method which utilizes latent variables to model response (target) variables from changes in the input (explanatory) variables. Two PLSR models were trained and tested using image features as input variables: (I) training and test data were randomly selected from the MFA-DT image data set to assess the impact of image-to-image variability on the PLSR prediction accuracy. (II) Training and test data consist of randomly selected independent data sets (training MFA-70-1 to MFA-90-5 excluding test data MFA-80-3 and MFA-90-5) and was used to explore the ability to translate a pretrained PLSR model to new crystallization experiments, therefore assessing the impact of variability between different, independent image data sets on the PLSR prediction accuracy. The MFA phase diagram in Figure 6 was used to calculate expected crystal suspension densities under isothermal conditions during stepwise heating for all image data sets (see example temperature profile Figure 1 P1). Image feature preprocessing was limited to discrete binning and averaging and subsequent mean-centering.

Figure 7 shows (top) parity plots and (bottom) residuals for both PLSR models. Using binned image features ($k_{av} = 5$), it is possible to predict crystal suspension densities with a mean residual error of 0.38 and 2.62 mg/mL assessed using test data (I, Figure 7A) from the same data set and (II, Figure 7B) from fully independent experiments, respectively. Image-to-image variability does not seem to have a significant impact on the PLSR prediction comparing performance metrics in Figure 7A using (light blue) no binning and (dark blue) a bin size of $k_{av} = 5$. In contrast, residuals of independent data sets (Figure 7B bottom, red and blue) show a nonrandom error distribution, indicating that some parts of the variability in the response cannot be accurately predicted using the pretrained PLSR model. This nonrandom noise could be attributed to small differences in the crystal suspension which are not fully captured such as its specific particle size and shape distribution which however have an impact on extracted image features. The number of latent variables is significantly reduced using a PLSR training data set with image feature data from multiple, independent data sets as shown in Figure 7B (ncmp = 9), indicating a lower risk for overfitting. Absolute PLSR weights (R) are highest for TENG, LAPV, RngGaussV157, IntV, and GLLV5 (for details, see Figure S4). TENG (Tenengrad function33) is a gradient magnitude-based method involving the Sobel operator. LAPV is the calculated image variance after the Laplacian operator is applied to determine second derivatives for passing high spatial frequencies associated with sharp edges. All other image features aim to quantify local (RngGaussV, GLLV) and global (IntV) image intensity variance. Interestingly, while the absolute gradient magnitude seems to be well captured using TENG, multiple image features are used to provide a good correlation between observed changes in the suspension density and the local or global image intensity variance. Differences in the quantification of the image intensity variance between the individual image features are mainly related to the type and the size of the convolution filter used during image transformation, which consequently both could be used as targets for future, application-specific feature engineering efforts. In the case of RngGaussV, a range filter is used during the image transformation. The impact of changing kernel sizes for range filters was discussed in the context of crystallization image characterization with example images provided in Figure 3. Overall, the pretrained PLSR model in Figure 7B is able to explain most physical changes in both independent data test sets (goodness of prediction, $Q^2 = 0.838$), and the relative prediction error is most significant for highly diluted crystal suspensions. While out of the scope for this study, additional information from object detection and characterization methods could be used to potentially improve crystal suspension density predictions for these diluted crystal suspensions.

Besides size and shape information, crystal suspension density predictions provide details on the system’s material balance to help inform population balance models during parameter estimation. Our results have shown for the first time that these can be derived from crystallization image data, significantly improving the scope of image data analysis applications for process characterization.

4. CONCLUSIONS

An image analysis framework for the routine characterization of crystallization image data was developed to extract a total of 80 image features from targeted image transformations including a parametrization of the pixel intensity histogram and other basic image intensity statistics. The developed methodologies were used to analyze images collected during small-scale crystallization experiments of MFA in changing solvent mixtures of DIG and WAT (70:30–90:10, w/w). Extracted image features were assessed for two important applications, A1 and A2, representative of key stages in crystallization process development. Specifically, the features were used for (A1) clear and cloud point detection to determine the MFA phase diagram and (A2) as input variables for a PLSR model to estimate crystal suspension densities up to 40 mg/mL.

(A1) For clear ($n = 12$) and cloud point ($n = 12$) detection, a simple threshold for the image feature signal HELM, was able to reliably detect both dissolution and onsets of crystal nucleation with a mean absolute error of 0.42 °C compared to a manually assessed, user-defined ground truth. For image feature-based clear point detection, the mean absolute error was even smaller with 0.19 °C or 0.42 g/mL. The image feature-based detection method significantly improves the determination of solubility and metastable zone width compared to the often employed transmission-based detection method which was unable to detect crystal suspension densities up to 15.26 mg/mL with a mean absolute error of 5.80 °C compared to the user-defined ground truth.
The implemented image analysis framework provides capabilities to automatically and accurately interrogate routinely collected image data from crystallization experiments. It is therefore significantly expanding the use of image analysis methodologies to complement traditionally employed methods focused on object detection. Such automated data analysis tools maximizing the useful information yield from small scale material sparing experiments have a significant role in enabling digital design approaches and the application of crystallization process modeling in early process development. Future efforts might focus on further exploring and developing targeted routines for image feature extraction of crystal suspension images to improve the accuracy and precision of pretrained PLSR models to independent crystallization image data sets. The image analysis framework also has great potential to be translated to other image-based systems and techniques for crystallization process characterization. Real-time image feature analysis could be applied for process control as a decision tool, reducing the need for more expensive, dedicated probes.

ASSOCIATED CONTENT

Supporting Information
The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.cgd.1c01118.

Direct image feature extraction and multivariate analysis for crystallization process characterization (PDF)

AUTHOR INFORMATION

Corresponding Author
Alastair J. Florence – Technology and Innovation Centre, EPSRC CMAC Future Manufacturing Research Hub, Glasgow G1 1RD, U.K.; Strathclyde Institute of Pharmacy & Biomedical Sciences, University of Strathclyde, Glasgow G4 0RE, U.K.; orcid.org/0000-0002-9706-8364; Email: alastair.florence@strath.ac.uk

Authors
Frederik J. S. Doerr – Technology and Innovation Centre, EPSRC CMAC Future Manufacturing Research Hub, Glasgow G1 1RD, U.K.; Strathclyde Institute of Pharmacy & Biomedical Sciences, University of Strathclyde, Glasgow G4 0RE, U.K.; orcid.org/0000-0001-5245-0503
Cameron J. Brown – Technology and Innovation Centre, EPSRC CMAC Future Manufacturing Research Hub, Glasgow G1 1RD, U.K.; Strathclyde Institute of Pharmacy & Biomedical Sciences, University of Strathclyde, Glasgow G4 0RE, U.K.; orcid.org/0000-0001-7091-1721

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.cgd.1c01118

Notes
The authors declare no competing financial interest.

Data underpinning this publication are openly available from the University of Strathclyde KnowledgeBase at https://doi.org/10.15129/75002e1c-80dc-42ab-984a-344f1e9e03

ACKNOWLEDGMENTS

The authors thank EPSRC and the EPSRC Future Continuous Manufacturing and Advanced Crystallization Research Hub (Grant ref: EP/P006965/1) for funding this work. The authors acknowledge that parts of this work were carried out in the CMAC National Facility supported by a UK Research Partnership Fund (UKRPF) award from the Higher Education Funding Council for England (HEFCE) (Grant ref HH13054).

REFERENCES

(1) Mullin, J. W. Crystalization; Elsevier, 2001.
(2) Brown, C. J.; et al. Enabling precision manufacturing of active pharmaceutical ingredients: workflow for seeded cooling continuous crystallisations. Molecular Systems Design & Engineering 2018, 3, 518–549.
(3) Urwin, S. J.; Levilain, G.; Marziano, I.; Merritt, J. M.; Houson, I.; Ter Horst, J. H. A Structured Approach to Cope with Impurities during Industrial Crystallization Development. Org. Process Res. Dev. 2020, 24, 1443–1456.
(4) Ottoboni, S.; Shahid, M.; Steven, C.; Coleman, S.; Meehan, E.; Barton, A.; Firth, P.; Sutherland, R.; Price, C. J. Developing a Batch Isolation Procedure and Running It in an Automated Semicontinuous Unit: AWI CFD25 Case Study. Org. Process Res. Dev. 2020, 24, 520–536.
(5) Prasad, E.; Robertson, J.; Halbert, G. W. Improving consistency for a Mefenamic acid immediate release formulation. J. Pharm. Sci. 2020, 109, 3462–3470.
(6) Guidance for Industry Guidance for Industry, PAT - A Framework for Innovative Pharmaceutical Development, Manufacturing, and Quality Assurance; FDA, 2004.
(7) Siddique, H.; Brown, C. J.; Houson, I.; Florence, A. J. Establishment of a Continuous Sonocrystallization Process for Lactose in an Oscillatory Baffled Crystallizer. Org. Process Res. Dev. 2015, 19, 1871–1881.
(8) McGinty, J.; Chong, M. W.; Manson, A.; Brown, C. J.; Nordon, A.; Sefcik, J. Effect of process conditions on particle size and shape in continuous antisolvent crystallisation of lovastatin. Crystals 2020, 10, 925.
(9) Agimelen, O. S.; Jawor-Baczynska, A.; McGinty, J.; Dziwiersz, J.; Tachtatzis, C.; Cleary, A.; Haley, I.; Michie, C.; Andonovic, I.; Sefcik, J.; Mulholland, A. J. Integration of in situ imaging and chord length distribution measurements for estimation of particle size and shape. Chem. Eng. Sci. 2016, 144, 87–100.
(10) Ferreira, C.; Cardona, J.; Agimelen, O.; Tachtatzis, C.; Andonovic, I.; Sefcik, J.; Chen, Y.-C. Quantification of particle size and concentration using in-line techniques and multivariate analysis. Powder Technol. 2020, 376, 1–11.
(11) Brown, C. J.; Ni, X.-W. Evaluation of Growth Kinetics of Antisolvent Crystallization of Paracetamol in an Oscillatory Baffled Crystallizer Utilizing Video Imaging. Cryst. Growth Des. 2011, 11, 3994–4000.
(12) Ochsnebein, D. R.; Vetter, T.; Schorsch, S.; Morari, M.; Mazzotti, M. Agglomeration of Needle-like Crystals in Suspension: I. Measurements. Cryst. Growth Des. 2015, 15, 1923–1933.
(13) Rajagopalan, A. K.; Schneeberger, J.; Salvatori, F.; Böttsch, S.; Ochsnebein, D. R.; Oswald, M. R.; Pellefey, M.; Mazzotti, M. A comprehensive shape analysis pipeline for stereoscopic measurements of particulate populations in suspension. Powder Technol. 2017, 321, 479–493.
(14) Cardona, J.; Ferreira, C.; McGinty, J.; Hamilton, A.; Agimelen, O. S.; Cleary, A.; Atkinson, R.; Michie, C.; Marshall, S.; Chen, Y.-C.;
Sefcik, J.; Andonovic, I.; Tachtatzis, C. Image analysis framework with focus evaluation for in situ characterization of particle size and shape attributes. *Chem. Eng. Sci.* **2018**, *191*, 208–231.

(15) Bruglia, M. L.; Sefcik, J.; Ter Horst, J. H. Measuring Secondary Nucleation through Single Crystal Seeding. *Cryst. Growth Des.* **2019**, *19*, 421–429.

(16) Doerr, F. J. S.; Florence, A. J. A micro-XRT image analysis and machine learning methodology for the characterization of multi-particle capsule formulations. *International Journal of Pharmaceutics: X* **2020**, *2*, 100041.

(17) Simon, L. L.; Abbou Oucherif, K.; Nagy, Z. K.; Hungerbuhler, K. Bulk video imaging based multivariate image analysis, process control chart and acoustic signal assisted nucleation detection. *Chem. Eng. Sci.* **2010**, *65*, 4983–4995.

(18) Tachtatzis, C.; Sheridan, R.; Michie, C.; Atkinson, R. C.; Cleary, A.; Dzwierz, J.; Andonovic, I.; Briggs, N. E. B.; Florence, A. J.; Sefcik, J. Image-based monitoring for early detection of fouling in crystallisation processes. *Chem. Eng. Sci.* **2015**, *133*, 82–90.

(19) Mullins, D.; Jones, E.; Glavin, M.; Coburn, D.; Hannon, L.; Clifford, E. A novel image processing-based system for turbidity measurement in domestic and industrial wastewater. *Water Sci. Technol.* **2018**, *77*, 1469–1482.

(20) Onyemelukwe, I. I.; Nagy, Z. K.; Rielly, C. D. Solid-liquid axial dispersion performance of a mesoscale continuous oscillatory flow crystalliser with smooth periodic constrictions using a non-invasive dual backlit imaging technique. *Chemical Engineering Journal* **2020**, *382*, 122862.

(21) Mitchell, N. A.; Frawley, P. J. Nucleation kinetics of paracetamol:methanol solutions from metastable zone widths. *J. Cryst. Growth* **2010**, *312*, 2740–2746.

(22) Jiang, S.; Ter Horst, J. H. Crystal nucleation rates from probability distributions of induction times. *Cryst. Growth Des.* **2011**, *11*, 256–261.

(23) Previtt, J. M. S. *Picture Processing and Psychopictorics*; Academic Press: New York, 1970; Vol. 10; pp 15–19.

(24) Clarke, M. R. B.; Duda, R. O.; Hart, P. E. Pattern Classification and Scene Analysis. *Journal of the Royal Statistical Society. Series A (General)* **1974**, *137*, 442.

(25) Brenner, J. F.; Dew, B. S.; Horton, J. B.; King, T.; Neurath, P. W.; Selles, W. D. An automated microscope for cytologic research a preliminary evaluation. *Journal of Histochemistry & Cytochemistry* **1976**, *24*, 100–111.

(26) Krotkov, E.; Martin, J.-P. Range from focus. *Proceedings. 1986 IEEE International Conference on Robotics and Automation*. 1986; pp 1093–1098.

(27) Meyer, Y. *Wavelets and Operators*; Cambridge University Press, 1992; Vol. 1.

(28) Eskicioglu, M. A.; Fisher, P. S. Image Quality Measures and Their Performance. *IEEE Transactions on Communications* **1995**, *43*, 2959–2965.

(29) Santos, A.; Ortiz De Solórzano, C.; Vaquero, J. J.; Peña, J. M.; Malpica, N.; Del Pozo, F. Evaluation of autofocus functions in molecular cytogenetic analysis. *J. Microsc.* **1997**, *188*, 264–272.

(30) Firestone, L.; Cook, K.; Culp, K.; Talsania, N.; Preston, K. Comparison of autofocus methods for automated microscopy. *Cytometry* **1991**, *12*, 195–206.

(31) Subbarao, M.; Choi, T.-S.; Nikzad, A. *Focusing Techniques. Machine Vision Applications, Architectures, and Systems Integration*, Vol. 1823. SPIE, 1992; pp 163–174. https://doi.org/10.1117/12.132073

(32) Geusebroek, J. M.; Cornelissen, F.; Smeulders, A. W.; Geerts, H. Robust autofocusing in microscopy. *Cytometry* **2000**, *39*, 1–9.

(33) Pech-Pacheco, J. L.; Cristóbal, G.; Chamorro-Martinez, J.; Fernández-Valdivia, J. Diatom autofocusing in brightfield microscopy: A comparative study. *Proceedings - International Conference on Pattern Recognition* **2000**, *15*, 314–317.

(34) Helmli, F. S.; Scherer, S. Adaptive shape from focus with an error estimation in light microscopy. ISPA 2001. *Proceedings of the 2nd International Symposium on Image and Signal Processing and Analysis*; ISPA, 2001; pp 188–193.

(35) de Jong, S. *SIMPLS: An alternative approach to partial least squares regression*. *Chemometrics and Intelligent Laboratory Systems* **1993**, *18*, 251–263.

(36) Liu, H.; Shah, S.; Jiang, W. On-line outlier detection and data cleaning. *Comput. Chem. Eng.* **2004**, *28*, 1635–1647.

(37) Abdul Mudalip, S. K.; Abu Bakar, M. R.; Jamal, P.; Adam, F. Solubility and dissolution thermodynamic data of mefenamic acid crystals in different classes of organic solvents. *J. Chem. Eng. Data* **2013**, *58*, 3447–3452.

(38) Reus, M. A.; Van Der Heijden, A. E.; Ter Horst, J. H. Solubility Determination from Clear Points upon Solvent Addition. *Org. Process Res. Dev.* **2015**, *19*, 1004–1011.

(39) Svoboda, V.; Machioniqbaile, P.; McGinty, J.; Connor, L. E.; Oswald, I. D.; Sefcik, J. Continuous Cocrystallization of Benzoic Acid and Isonicotinamide by Mixing-Induced Supersaturation: Exploring Opportunities between Reactive and Antisolvent Crystallization Concepts. *Cryst. Growth Des.* **2017**, *17*, 1902–1909.