Batch to Batch Variation Identification in Production of Crystalline Products Based on Image Analysis and Particle Shape Classification

Jing J. Liu\(^a\), Jian Wan\(^b\), Caiyun Ma\(^b\), Yangdong Hu\(^a\) and Xue Z. Wang\(^b\)*

\(^a\)School of Chemical Engineering, Ocean University of China, Shangdong, PR China
\(^b\)Institute of Particle Science and Engineering, University of Leeds, Leeds LS2 9JT, UK.

*Correspondence author: Xue Z. Wang, x.z.wang@leeds.ac.uk

Abstract
For crystalline particulate products such as pharmaceuticals, shape, as a product quality variable, can be as important as size. This paper presents a methodology to derive shape descriptors from microscope images of particulate products and the use of the descriptors for classification and identification of batch to batch variations. Data from batch runs of laboratorial and industrial crystallisers are analysed.

Keywords: batch to batch variation, image analysis, shape descriptors, principal component analysis, classification, process analytical technology

1. Introduction
Quality assurance in particulate product manufacture is moving towards using on-line, in-line and at-line process analytical technology (PAT) for rapid and detailed characterization of chemical and physical properties of samples. One of the properties is shape, because it can affect product performance properties such as bioavailability of drug particles and particles’ processability so size alone as often defined as volume equivalent diameter is over simplified and can be misleading for certain applications. Optical microscopy has been proved to be one of the most effective techniques for determining particle shape and size off-line or on-line [1]. The images can be analyzed using image analysis algorithms [2,3] to obtain particle shape information.

In industry, shape descriptors that have some physical meanings are often used, such as maximum length, aspect ratio and roundness. They are simple and intuitive, but can cause loss of the original morphological information since the original shape cannot be reconstructed from them. In this work, principal component analysis was investigated for processing a set of new shape descriptors calculated from the combination of different physical size measurements of the particle contours. The principal components are then used for classification with the aim of identification of batch to batch variations in quality. The results are compared with Fourier descriptors.

2. Shape descriptors
An automated imaging system for analysis of the size and shape of particles, the PharmaVision 830 of Malvern Instruments Ltd was used to measure the physical size dimensions of particles. The samples are homogeneously dispersed onto a glass slide using a special sample preparation device based on pressurized air. The glass slide is placed under a CCD camera that scans the sample for particle image acquisition. Particles are segmented from the frames by embedded computer software obtaining the individual particle images and a variety of particle descriptors.
Basic shape measurements from the PVS 830 contain three descriptors, namely, Roundness, Contour/Area and Convexity, whilst the majority of other measurements are associated to size, i.e., Mean Diameter, Diameter, Length, Width, Maximum Distance, Area and Volume. Given that the particle size can be expressed in various ways according to the morphological dimensions of the particles such as length, breadth, volume or maximum distance, a large set of physical shape descriptors with various orders can be created by the arithmetic combination of the size measurements. Each of such combinations is sensitive to a specific attribute of shape according to the size measurements selected. Based on this principle, the size measurements obtained by the PVS 830 were combined as follows to create 85 new shape descriptors [4]:

\[
\begin{align*}
&\{ \frac{V}{L^3}, \frac{A}{L^2}, \frac{L_1}{L_2}, \frac{L_1 L_2}{L_3}, \frac{V}{L_1 L_2 L_3}, \frac{A}{L_1 L_2 L_3} \} \\
&\end{align*}
\]

where \( V \) represents the particle volume calculated from the area and mean diameter, \( A \) is the particle area and \( L \) represents the different particle size measurement, i.e., Length, Width, Diameter, Mean Diameter and Maximum Distance. In combination with Roundness, Contour/Area and Convexity, a total of 88 descriptors for each particle.

For the image of each particle, the 88 descriptors were calculated. Then for all the particles, principal component analysis (PCA) was performed, and the selected principal components (PCs) were used in the next step for classification. Since all the 88 descriptors have some sort of physical meaning, contribution plots can be used to calculate the contribution of each descriptor to the selected PCs, and so produce a ranking of them with regard to their contribution to the classification results. The contribution of every original descriptor to two clusters is estimated based on the original data and the classification result, which is calculated by the following equation:

\[
\varphi_i = \frac{C_i^1 - C_i^2}{\max_j d_{ij}}, \quad i = 1, 2, \ldots, n
\]

where \( \varphi_i \) is the estimated contribution of the \( i \) th physical descriptor; \( C_i^1 \) and \( C_i^2 \) are the mean values of the \( i \) th physical descriptor for the first classified cluster and the second classified cluster, respectively; \( d_{ij} \) is the \( i \) th physical descriptor for the \( j \) th particle in the sample while \( j \) is the index of the individual particle in the sample. The estimated contribution can be intuitively interpreted as the descriptor contributes most to the resulting classification. The calculated values of \( \varphi_i \) can be normalized to obtain a contribution fraction of each variable in the dataset.

A further comparison for shape classification was carried out using latent Fourier descriptors. Fourier descriptors for particle shape have been the subject of several previous studies [5]. Unlike physical shape factors, Fourier descriptors can be considered as a kind of latent descriptors, i.e., they do not provide a direct physical meaning of the described object although they do provide means to retain the morphological information and the capability to reconstruct the shape. In order to obtain the descriptors, a chain code was applied to obtain the coordinates of the points \((x, y)\) composing the shape boundary to derive the shape signatures. In previous studies on classification of particles produced in crystallization processes, we have explained the details of the procedure to calculate the latent Fourier descriptors from the contour coordinate points [6].
Image analysis for particle shape characterisation

3. Classification
Classification techniques of Kohonen neural networks, Adaptive Resonance Theory 2 (ART2) and fuzzy C-means were used for classification of particles based on the shape descriptors. The results by all three methods are consistent and comparable, so here only the result of fuzzy C-means clustering is presented. As a method of clustering which allows one piece of data to belong to two or more clusters, fuzzy C-means is based on minimization of the following objective function:

$$J_m = \sum_{i=1}^{N} \sum_{j=1}^{C} u_{ij}^m \| x_i - c_j \|^2, \quad m \geq 1,$$

where \( m \) is any real number greater than 1, \( u_{ij} \) is the degree of membership of \( x_i \) in the cluster \( j \), \( x_i \) is the \( i \)th of \( d \)-dimensional measured data, \( c_j \) is the \( d \)-dimension centre of the cluster, and \( \| \cdot \| \) is any norm expressing the similarity between any measured data and the centre. Fuzzy partitioning is carried out through an iterative optimization of the objective function with the update of membership \( u_{ij} \) and the cluster centres \( c_j \) by:

$$u_{ij} = \left( \sum_{k=1}^{C} \frac{\| x_i - c_j \|}{\| x_i - c_k \|} \right)^{-2}, \quad c_j = \frac{\sum_{i=1}^{N} u_{ij}^m x_i}{\sum_{i=1}^{N} u_{ij}^m}.$$

This iteration is to stop when \( \max_j \left\| u_{ij}^{(k+1)} - u_{ij}^{(k)} \right\| < \varepsilon \), where \( \varepsilon \) is a termination criterion between 0 and 1, whereas \( k \) are the iteration steps. This procedure converges to a local minimum or a saddle point of the object function.

4. Case studies
4.1. Case study 1
Case study 1 is concerned with classification of the 55 crystalline particles shown in Figure 1. The purpose is to compare the results of classification of using different descriptors. For each of the particle image, three types of descriptors were calculated: 64 Fourier descriptors; 88 shape descriptors according to expression (1) plus roundness, contour/area and convexity; and three principal components obtained from PCA processing of the 88 descriptors. Table 1 gives the classification results using fuzzy C-means. Other two classification techniques, Kohonen and adaptive resonance theory 2 – ART2 gave very similar results as fuzzy c-means. Examination of Table 1, it can be concluded that the classification results based on the three types of descriptors are very consistent. The best number of classes is two. Particles classified into Class 1 are more like in prismatic shape e.g. particle number 3, while particles assigned to Class 2 have larger aspect ratio e.g. particle number 55. Particles that are not consistent in class assignment between the three types of descriptors are those having more complicated shape than either prismatic or needle. Fourier descriptors are considered as being able to preserve more original shape information, but they don’t have physical meanings. Although 88 descriptors and three PCs obtained gave similar results, we still feel that PCA processing should be used since it removes dependencies between 88 descriptors. Therefore in the next two case studies, PCA processed descriptors are used.
Table 1. Comparison of classification results for the 55 particles using three types of descriptors

<table>
<thead>
<tr>
<th>Shapes</th>
<th>64 Fourier descriptors</th>
<th>Normalized 88 shape descriptors</th>
<th>Three PCs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class 1: Prismatic</td>
<td>33 members: 1-21,23,26,28-30,35,36,38,40,45,48,52</td>
<td>25 members: 1-5,7-21,29,40,46,48,52</td>
<td>27 members: 1-5,7-23,29,40,46,48,52</td>
</tr>
</tbody>
</table>

4.2. Case study 2: application to a laboratory process for crystallization

When crystallizing from aqueous solutions, L-glutamic acid is known to have two polymorphic forms, α and β. This case study examines differences in morphology of crystals obtained in three batch runs. Samples from two batches were labeled as LGA A01 and LGA A02. The two batch runs used the same cooling strategy, i.e. cooled from 70 °C to 10 °C using a linear cooling rate of 0.5 °C/min. A third sample, LGAAB001, was obtained from a batch run using the same reactor, and also cooled from 70 °C to 10 °C but using a different cooling rate of 0.2 °C/min.

For each experiment, crystals obtained were filtered and dried. The dried particles were characterized using the imaging instrument. Figure 2 shows example frames for each sample. For each segmented particle, 88 descriptors were calculated, and then processed using PCA. Classification result is shown in Figure 3. It can be seen that samples 1 and 2 are overlapped, while sample 3 particles are grouped around a different cluster centre.
Image analysis for particle shape characterisation

Figure 2 Sample frames of the three sets: (a) LGA A01, (b) LGA A02 containing α-form particles, and (c) LGAAB001 containing both α-form and β-form of L-glutamic acid

Contribution analysis found that the roundness descriptor provides the largest difference in shape between the particles of the batches. Figure 4 shows the distribution of roundness for particles of the three batches.

![Figure 4. Distribution of roundness for particles of the three batches.](image)

4.3. Case study 3 Application to an industrial crystallization process

This example compares the samples obtained from two different crystallisation batches of L-glutamic acid produced in an industrial-scale pilot plant at Switzerland. The two batches were carried out in a 250 L batch reactor under supersaturation control, one batch with supersaturation, $S = 1.1$, and the other with $S = 1.2$. The samples were obtained towards the end of the runs from the reactor using a long sample stick to scoop up the crystals from the solution. The crystals were filtered and dried. The two batches produced the stable polymorph, β, hence the characteristic needle shape of the polymorph was observed in both runs. Nevertheless, it is of interest to know the morphological differences between the needles as a consequence of the supersaturation effect on the growth of the crystal faces, that may be difficult to detect visually from micrographs of the samples.

Images of the two samples were analysed. The samples of $S = 1.1$, lgaS1101, contained 7511 particles measured. The sample of $S = 1.2$, lgaS12, contained 20000 particles measured and, for the analysis of appropriate descriptors, this sample was subdivided into three sub-samples in order to make similar the number of particles between the samples: the new samples being lgaS1201 (6500 particles), lgaS1202 (6500 particles) and lgaS1203 (7000 particles).

The four sample datasets were analysed for descriptor identification. The output indicated that the main difference in particle shape between the two batches lies in the Roundness of the crystals. The supersaturation of $S = 1.1$ led to the production of longer needles, reflected as lower values of roundness. Figure 5 displays the distribution of
roundness from the four sample datasets analysed, where it is clearly observed that the single sample distribution for $S = 1.2$ corresponds to particles with lower Roundness compared to the fairly similar three sample datasets for $S = 1.1$.

![Figure 4](image1.png)  
**Figure 4** The distribution of the shape descriptor, roundness, among three batches of case study 2

![Figure 5](image2.png)  
**Figure 5** The distribution of the shape descriptor, roundness, among the four batches of case study 3

5. Conclusion
A methodology for identifying batch to batch variation in terms of shape difference in manufacture of particulate products is presented. The method is based on the derivation of new shape descriptors produced from multiple combinations of physical size measurements obtained by imaging and image analysis. The new descriptors were processed using principal component analysis, prior clustering techniques are used. Three case studies with varied complexity were used to illustrate the methodology.

6. Acknowledgements
Supports from UK Engineering and Physical Sciences Research Council (EPSRC grant references: EP/E045707/1 and EP/H008012/1), and from Malvern Instruments Ltd (for IntelliSense and Vision projects) are greatly acknowledged.

References