Journal of
Applied Crystallography

ISSN 1600-5767

Received 29 July 2013
Accepted 4 November 2013

Cluster method for analysing surface X-ray diffraction data sets using area detectors

Steven J. Leake,* Mathilde L. Reinle-Schmitt, Irakli Kalichava, Stephan A. Pauli and Philip R. Willmott

Swiss Light Source, Paul Scherrer Institut, CH-5232 Villigen, Switzerland. Correspondence e-mail: steven.leake@psi.ch

An automated cluster algorithm is described, applicable to any image where a signal is to be analysed. The algorithm is employed in the context of surface X-ray diffraction data and extended to automate the data reduction process, which at present limits both the lead time to and the reliability of the retrieved structural information. A detailed evaluation of the constraints used to automate surface X-ray diffraction data analysis is provided. To overcome limitations of the algorithm and the experiment itself in certain geometries, the full field of view of area detectors is exploited to obtain orders of magnitude improvements in data collection. The method extends the surface X-ray diffraction technique to new systems and highlights the often archaic approach to the analysis of data collected with a two-dimensional detector.

1. Introduction

The increased brightness of X-ray sources has led to the rapid development of two-dimensional pixel detectors (Brönnimann *et al.*, 2003; Team, 2010), resulting in a significant increase in data volume. The potential of the additional information measured (Mariager *et al.*, 2009; Schlepütz *et al.*, 2011) is often overlooked, limiting the impact of the science and the availability of beamtime to others, and at significant cost to the facilities themselves. Here, we demonstrate an automated approach to data extraction, coined the 'cluster method', and several methods to improve data analysis related to surface X-ray diffraction (SXRD).

SXRD has seen a renaissance in the past decade as reliable structure factors can now be measured with a single exposure of a two-dimensional detector (Schlepütz *et al.*, 2005). The ever improving sources, endstations and collection regimes, such as trajectory scanning, generate ever increasing volumes of data (Rivers, 2007). At present, intensities (the measured intensity being the modulus of the structure factor squared, $I = |F|^2$) with sufficient statistics to reconstruct the electron density of the diffracting surface can be acquired in less than 24 h. The main bottleneck of SXRD begins at the structure-factor-extraction stage. Present extraction methods are entirely manual and are highly labour intensive. The subsequent phasing of the data is also problematic, but will not be covered here (Bjoerck *et al.*, 2008; Yacoby *et al.*, 2003; Saldin *et al.*, 2001). It is common for the lead time from measurement to atomic structure to be of the order of a year or more. Errors associated with the extraction of the measured data propagate through the analysis; therefore their minimization is critical to its success. Automation of the extraction process brings two distinct advantages. Firstly, as the number of measured structure factors moves towards the tens of thousands as more data-hungry techniques are developed (Pauli *et al.*, 2012), the accumulation of human error complicates or even voids the subsequent analysis. Secondly, automation has the potential to produce reliable structure factors in real time, improving experiment efficiency, significantly reducing the associated lead time to results and increasing the reliability of the obtained data.

Several automated signal identification methods for diffraction peaks are presented in the literature. In the simplest case, a box is drawn around the predicted location of the peak and some combination of the pixels on the perimeter of the box is used to estimate the background. However, in practice the peak shapes on the detector are rarely uniform, and thus the box includes background pixels in the integration, increasing statistical errors, particularly for weak reflections. Two different approaches have resulted: firstly, *a priori* information about the peak shape is calculated theoretically or learned from other peaks (Ford, 1974; Kabsch, 1988; Rossmann, 1979; Roth & Lewit-Bentley, 1982; Schoenborn, 1983; Lehmann & Larsen, 1974; Wilkinson *et al.*, 1988), and secondly, an objective approach is taken, where based on a background approximation the peak is defined by an intensity threshold (Sjölin & Wlodawer, 1981; Wlodawer & Sjölin, 1982; Filhol *et al.*, 1983). A move towards individual peak optimization was made with Gaussian curve fitting to pixel–intensity histograms (Spencer & Kossiakoff, 1980) and furthered by Bolotovsky *et al.* (1995), who applied a statistical approach whereby a 'seed' at some pre-defined position, set by the crystal orientation matrix, is slowly expanded until the 'skewness' of the background pixel–intensity distribution is zero, *i.e.* a Poisson-like distribution. Since then, to our knowledge, no further generic methods for signal identification have been proposed in the diffraction field.

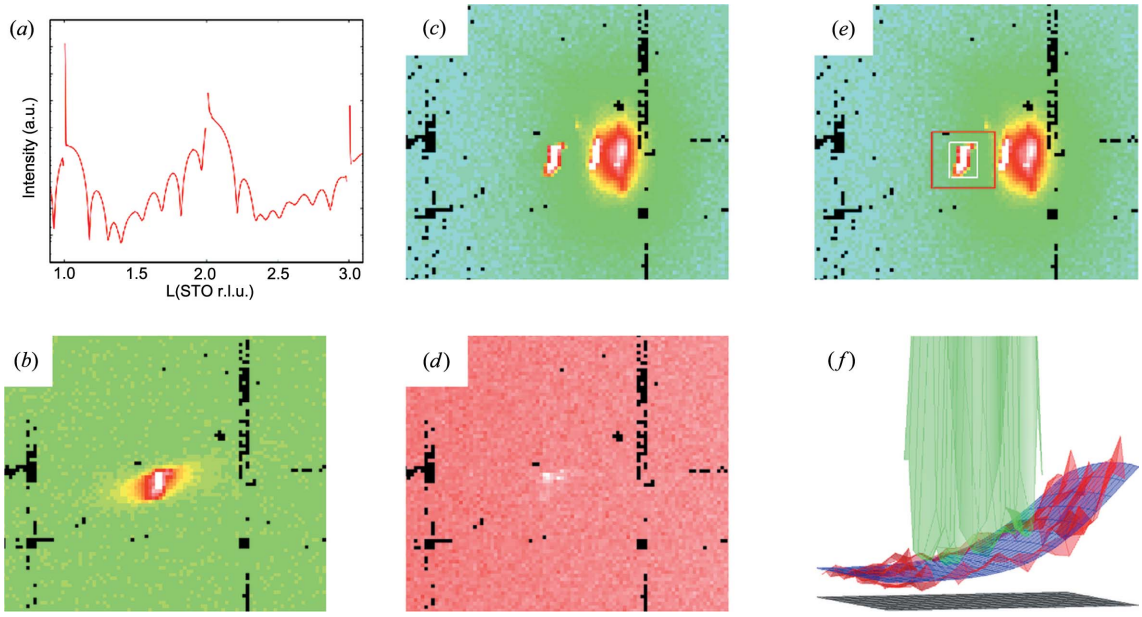

Figure 1
Examples of diffraction signal observed at different regions of (*a*) a CTR: (*b*) far from a Bragg peak, (*c*) close to a Bragg peak, (*d*) very weak signal, (*e*) signal region of interest (sROI, white box) and background region of interest (bROI, red box) of (*c*), and (*f*) surface plots of sROI and bROI. (*b*)–(*e*) have a logarithmic scale, scaled to the maximum of the signal, to enhance the background features. The black pixels were deemed unreliable after flatfield correction.

This article will provide an overview of the problems associated with SXRD signal analysis, describe the 'cluster method' and detail its application in an automated approach. A further discussion will outline a method to exploit the full two-dimensional detector image using a traditional scanning mode.

2. Background

2.1. Crystal truncation rods

A typical SXRD measurement samples the continuous distribution of intensity in reciprocal space found between Bragg peaks perpendicular to the surface, the so-called crystal truncation rods (CTRs). The shape of the signal on the detector corresponds to a convolution of the shape of the incident X-ray beam on the sample and the surface quality (mosaicity, defects, miscuts *etc.*). By rotating the detector the footprint's orientation can be held constant, although its size can vary (Schlepütz *et al.*, 2011). This continuity between consecutive data points, assuming a large oversampling ratio, paves the way for an automated approach by tracking the signal, which changes in shape in a slow and predictable way. A typical CTR is shown in Fig. 1(*a*). The measured signal consists of the structure factor associated with a crystalline sample plus a diffuse background due to defects, thermal diffuse scattering or sometimes spurious diffraction sources, such as the bulk substrate, beamline components or sample environment. At certain l positions (reciprocal lattice units, r.l.u.) along the rod, one observes the diffuse background associated with defects in the sample (Fig. 1*b*); the background generated by the broad tails of the substrate Bragg peak (Fig. 1*c*); and the very weak signal observed on the rod at the minima between two Laue oscillations (*d*). An example extraction of image (*c*) is shown in Fig. 1(*e*); the fitted background is shown in Fig. 1(*f*) as a blue two-dimensional surface.

In addition to a typical CTR signal we highlight the need to treat 'low-l' data with a different set of parameters, as the intersection of the Ewald sphere here with the CTR leads to an elongation of the footprint (Vlieg, 1997), at which point the open-slit geometry breaks down and a rocking scan measurement approach must be employed. The term 'low l' is energy and sampling-period dependent. Typically we approach any scatter below $l = 0.5$ in this manner, even for high-quality metal-oxide crystals. This approach is often time consuming and therefore not followed. However, a simple adjustment of the detector orientation and the extraction method can be used to obtain reliable data, should detector rotation be available.

2.2. Diffraction signals

Three main factors contribute to the size and shape of the signal and the background observed on a two-dimensional detector: the beamline, the sample and its environment.

Beamlines are highly tunable, and the footprint of the beam on the sample can be tailored to the experiment. In the ideal case, the diffracted signal is spread across as many pixels as possible to improve the efficiency of measurements, decrease the dependency on a few illuminated pixels and gain a better understanding of anomalous signals by improving the resolution of the diffraction features.[1]

[1] SXRD experiments are operated on both undulator and wiggler beamlines; the optics employed after the source tune the size of the illuminating beam which, coupled with the geometry, defines the size of the signal on the detector.

The physical size of the portion of the sample responsible for a diffraction peak determines its shape (reciprocal) and distribution (strain). The quality of the surface is paramount: mosaicity smears the signal out; defects in the substrate contribute to the background; and miscuts of the surface with respect to the nominal crystallographic plane lead to splitting of the CTR signal (Munkholm & Brennan, 1999).

SXRD measurements are normally made in vacuum to minimize radiation damage of the surface and/or interface region. Such measurements are usually carried out under a beryllium dome, which, although a weak scatterer, contributes significantly to the background in the form of an overall diffuse signal. Debye–Scherrer rings originating from the beryllium dome or its mount intersect intermittently with measured signals; at anti-Bragg positions they can be brighter than the signal being sought. The increased brightness available from undulator sources reduces powder rings to textured scatter from individual crystallites, making them harder to subtract. A solution implemented at many synchrotrons is a Kapton dome with a helium atmosphere, typically improving ∼5% of the measured data points. The drawback of this method is that Kossel lines (Kossel *et al.*, 1935) are now measurable, and these are significantly harder to fit as they cannot be approximated by a two-dimensional line of intensity with a Gaussian profile across its width when many Kossel lines are present. Both options require consideration when applying an automated approach. Fluorescence can also contribute to the background, depending on the detector type and any illuminated material with a line of sight to the detector itself.

2.3. Analysing signals

The approach used to analyse data manually has been detailed comprehensively elsewhere (Schlepütz, 2009), including the MATLAB (MathWorks, 2012) graphical user interface (GUI) used in the analysis, named *Scananalysis*. Two regions representative of the signal and the background are selected; a fit to the background then allows the structure factor to be extracted. Depending on the number of structure factors to extract this process can take several weeks and is often completed independently multiple times to test consistency. Some examples of signal that is difficult to analyse are shown in Fig. 2. The more complicated the shape of the signal, the more time consuming it is to extract the data, and the more likely the user's interpretation of the signal boundary will vary. Each data point is assigned a measure of quality, either 'g'ood, 'b'ad or 'd'ubious, a terminology which will be employed in the automation approach to identify points 'd' that need to be revised manually *a posteriori*. Traditional signal-to-noise ratio methods used in standard signal analysis, such as the Rose criterion, do not apply here, as the signal is spatially non-uniform. Our approach instead applies known physical constraints to identify the signal and continuity between consecutive data points to guide the algorithm.

2.4. Rocking curves

Prior to the advent of two-dimensional detectors, structure factors were obtained with a point detector using the rocking curve (RC) approach. Sample rotation around the axis perpendicular to the surface (denoted ω_y here; Schlepütz *et al.*,

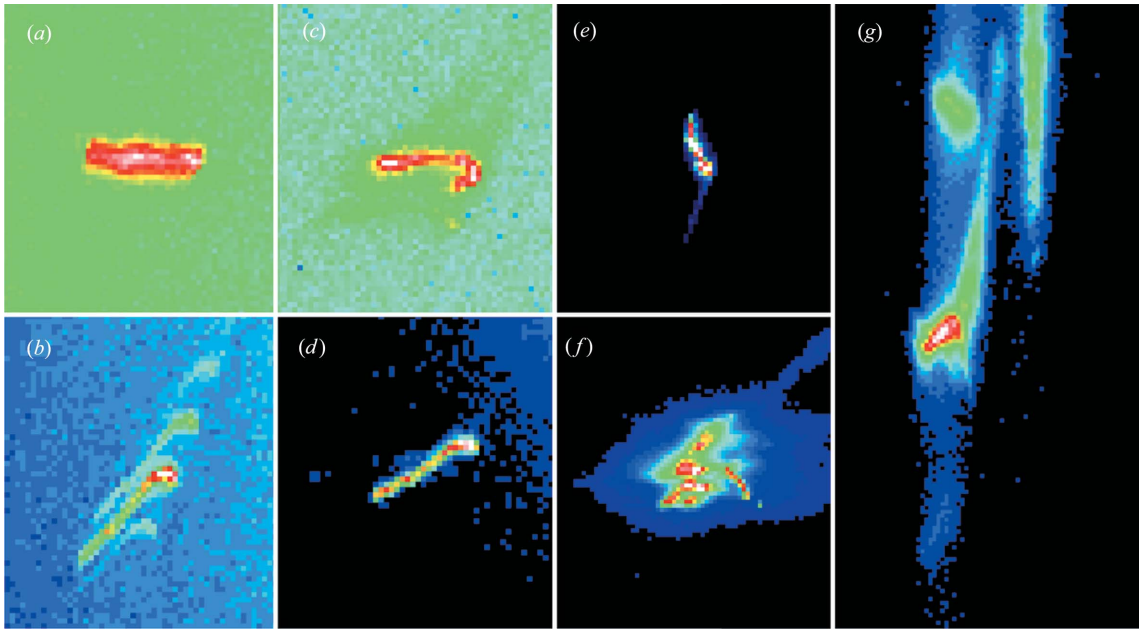

Figure 2
Examples of CTR signals that are problematic to extract, obtained from $LaNiO_3$ grown on $SrTiO_3$ substrates, and an ideal signal for comparison ($LaAlO_3$ film) (*a*): (*b*) split signal due to the miscut, (*c*) asymmetric signal ($LaAlO_3$ film), (*d*) bright spots at one end of the footprint, (*e*) tails on the footprint, (*f*) Bragg peak due to multiple domains of tetragonal $SrTiO_3$ below 110 K, and (*g*) multiple Laue oscillations intersecting the Ewald sphere at low *l*.

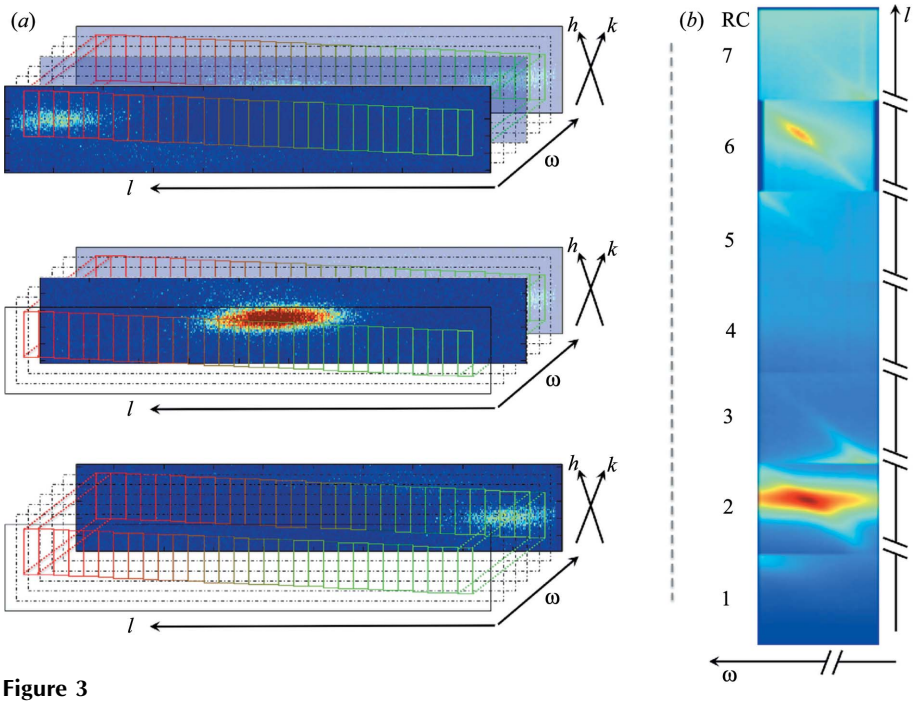

Figure 3
(*a*) An example of the geometry of the beginning, middle and end of a rocking curve (RC) carried out with a two-dimensional detector by rotating ω_y. The signal is split along the *l* direction, into the desired δl step, shown by red–green cuboids, and treated individually to produce multiple extracted RCs from a single scan. (*b*) One hundred and fifty extracted RCs, produced *via* the method detailed for (*a*) from seven RCs measured with an area detector, taken at strategic *l* values along a single superstructure crystal truncation rod. Note the two axes are broken to aid visualization of overlap between successive detector RCs.

2011) rotates reciprocal space such that the stationary point detector samples perpendicular to the CTR. Subsequent background subtraction and integration provide the structure factor.

One can think of each pixel on a two-dimensional detector as being a point detector. Thus, alignment of the long axis of the detector parallel to the *l* axis (Schlepütz *et al.*, 2011) means that a subsequent RC through a CTR produces a signal on the detector which traverses the long axis of the detector as is shown in Fig. 3(*a*). This is the CTR sweeping through the Ewald sphere as reciprocal space is rotated.

Segmentation of the stack of detector images into δl portions, *i.e.* along the detector's long axis, provides multiple extracted RCs from a single RC scan, which can be fitted, background subtracted and integrated to obtain the structure factors. Each δl portion is equivalent to a point detector, and thus the area detector in a single scan measures significantly more information than is normally used. Application of this method allows one to measure weak and diffuse signals with high accuracy and relatively quickly.

In principle an entire CTR could be measured using this method with enhanced resolution, as shown in Fig. 3(*b*), from seven RCs, measured at different *l* values. A total of 150 RCs were extracted, and their profiles are plotted as a function of *l*. In principle, if the detector rotation were not available, given the visible overlay between consecutive rocking curves, a three-dimensional interpolation approach could be employed to integrate such a data set, but this is beyond the scope of the present article.

3. Methods

3.1. Cluster method

A cluster method has been developed to extract complicated diffraction signal shapes. On the basis of some criterion, for example a threshold intensity, an automatic appraisal of the shape of the diffraction signal in the image is made; this is then selected and expanded using a convolution-type operation to generate a background. The output is a MATLAB file suitable for loading into the existing *Scananalysis* program used for manual extraction for checking 'd' or 'b' data points. *Scananalysis* itself was upgraded to integrate the new cluster method and is freely available upon request. The cluster method has also been written in Python.

The cluster tool has two modes of operation: connected signal is defined as either (i) those pixels connected along their common sides (*i.e.* above/below/left/right) or (ii) all pixels surrounding an individual pixel including pixels connected by corners. Similar arguments have been used to count photons in charge-coupled devices, employing only the first definition, called the 'droplet algorithm' (Livet *et al.*, 2000).

In order to identify all clusters above a defined threshold in a computationally friendly manner, the two-dimensional array is searched line by line. Once a pixel above the threshold is

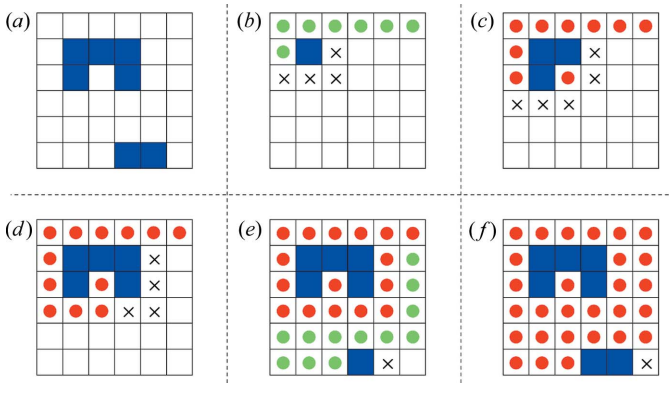

Figure 4
(*a*) Two-dimensional data containing two clusters to locate. (*b*)–(*f*) Graphical demonstration of consecutive steps the algorithm employs to identify the cluster. Red circles have been checked and proven to be below the threshold, green circles are checked in the current iteration, and crosses have to be checked in the next step as a result of finding a pixel with intensity above the threshold.

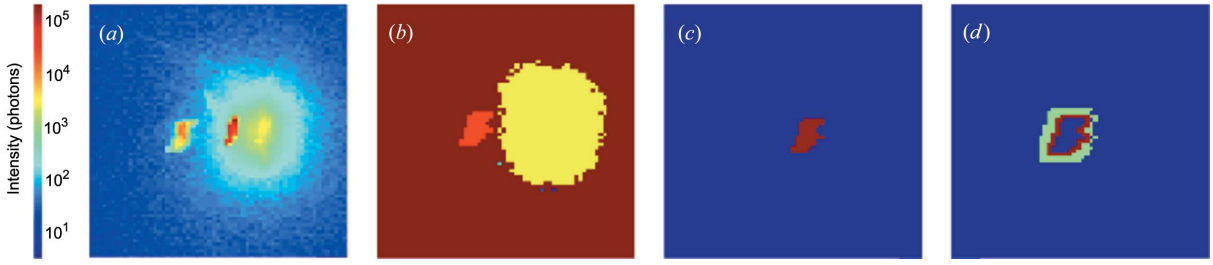

Figure 5
Example of the cluster procedure on the image from Fig. 1(*c*). (*a*) The signal to analyse, (*b*) clusters identified based on a threshold of 0.1%, (*c*) the identified signal cluster, and (*d*) the expanded signal and chosen background (light green) overlaying the original cluster (opaque).

identified, all its connecting pixels that are also above the threshold are determined, and then their adjacent pixels above the threshold, and so on, until no new adjacent pixels above the threshold are found. This is a single-pass operation. A single mask of clusters results, with each cluster assigned a number; this can be employed later to define connected clusters, such as in the case of sample miscuts, or remove erroneous scatter from the background results. Fig. 4 graphically demonstrates the cluster operation in its basic form on a 6×6 pixel array.

3.2. Definition of a threshold

In order to automate the cluster approach to traverse an entire CTR, an intensity threshold is required to identify the signal. The large intensity variation present along a CTR requires a dynamic threshold, with the initial value set depending on the type of sample (*i.e.* terminated bulk crystal or thin film) or the position in reciprocal space; the signal-to-noise ratio in weak regions can be <1 but in bright regions $>10^4$. Setting the initial guess of the threshold is sample and experimental-setup dependent; for an intense data point, *i.e.* close to a Bragg peak, 5–10% of the maximum intensity was sufficient. It is important to note that the cluster identification only needs to provide a rough signal shape; further convolution-type[2] operations and background fitting precisely define the structure factor.

3.3. Automation

In order to translate the cluster method and a signal threshold into an automated procedure sufficiently robust to overcome all the spurious signals one might encounter, several steps need to be taken. The first question to address is, what is the physical position of the signal on the detector? The experimental setup places the diffraction peak at the centre of the detector. The central pixel is thus the reference point. However, this point of reference is not guaranteed; in order to account for asymmetric intensity distributions of the signal (see Fig. 2), overlapping of Debye–Scherrer rings or, on rare occasions, the failure of a motor movement, a symmetric region around the central pixel, called the region of interest (ROI), is defined as that where the signal could possibly exist. The ulterior motive for defining this ROI is the speed improvement gained from analysing smaller images; in the case of a Pilatus detector the field of view is so large that less than 10% of it is required as the ROI, and often <1% corresponds to both the signal and its background. One could envisage a simple calculation of the reciprocal size of the beam footprint on the sample surface, but this constraint was tested and found to be often too strong as sample imperfections are inevitable across surfaces up to 10 mm in size. The solution was to use a larger ROI, search it for possible signal candidates using the cluster method and define the cluster closest to the central pixel as the signal.

The second question is, where is the boundary between signal and background? The dynamic threshold finds the signal itself within 1–2 pixels of the boundary: we define this as the signal region of interest (sROI). The background signal tends to extend beyond this in at least one dimension in the detector plane (see Fig. 2), and thus a background region of interest (bROI) is defined by a convolution of the sROI with an ($n \times m$) pixel array (nmROI), where n and m are integers defined by the analyst, minus the sROI and any previously identified clusters of spurious intensity such as intense Bragg peak tails, for example. A convolution approach was applied to aid asymmetric signal analysis. Here, if a more uniform bROI, for example a square, were used, after the removal of the sROI the distribution of pixels would be skewed across bROI, thus artificially weighting the fit in the more populated regions. The process is demonstrated in Fig. 5 on the image shown in Fig. 1.

Additional gains are achieved by invoking information learned from the previously extracted image, as CTRs are slowly varying when oversampled by a factor of ten. This is discussed further later. These methods increase computation time considerably and will be discussed only within the context of achieving reliable online structure factors at a beamline 'live'. Two approaches will be discussed. The 'basic' version works in the *Scananalysis* GUI but because of the visualization is significantly slower. Thus a command-line version with the same operation was developed for speed and to implement more computationally complex constraints in a flexible way; this is referred to as the 'advanced' approach.

[2] A true convolution approach was not employed in order to enhance computation speed, although the approach is equivalent to convolving with a two-dimensional Gaussian function.

3.4. SXRD data reduction: basic

A detailed description of the implementation of the cluster method to an SXRD data set follows. After the steps identified in the 'automation' section, the sROI is identified. The bROI is then fitted with a function – either a constant background, a Gaussian, a linear gradient, or a linear + Gaussian – and subtracted from the sROI, which is subsequently integrated. One of these methods is chosen for the entire data set. We know that in the vicinity of the Bragg peak it is likely that we will need 'linear + Gaussian' to fit the tail of the Bragg peak, whereas at the anti-Bragg position a 'constant' background is employed to prevent overfitting the noise. However, the sample type, quality and environment lead to spurious diffraction, and so this method will not work for all data points. The objective approach is to treat all of the data in the same way, and then a few data points can be reanalysed where necessary.

Several safeguards are in place to prevent the sROI ballooning out of control. A user-defined background is set whereby any reflection signal below this minimum threshold, typically 50–100 photon counts per pixel per second depending on scattering from the sample environment, is marked 'd'ubious and the sROI of the previous data point (prev_sROI) is invoked for the current point. This leads to several regions of dubious data points which need subsequent manual confirmation or adjustment. Along the length of the CTR the extent of the signal on the detector varies only very slowly; thus if the number of pixels defining the new signal is over 150% of that of the previous signal, the threshold is deemed too low and is adjusted to generate a cluster of approximately the same size as the previous image. The same notion is applied to the lower limit to prevent the cluster shrinking to zero. Assuming a perfect sample alignment, the central pixel should sit in the sROI. This can be relaxed to an overlap between consecutive sROIs if required, as slight misalignments can cause the signal to drift on the detector: for example, as one traverses a rod the signal can migrate slowly owing to the miscut of the surface.

The order in which the data are treated is also taken into consideration. Typically the algorithm is most challenged at anti-Bragg positions, where the signal is weakest. Therefore it can be advantageous to approach the data beginning near the Bragg peaks, where the signal is strongest, and work towards the anti-Bragg positions armed with the information gained, instead of iterating up the rod point by point.

In addition to the criteria mentioned previously, we automatically avoid data points within $0.01l$ of the Bragg peaks, as they are often swamped by the Bragg peak itself. Any data points that do not meet the defined criteria are marked 'd'ubious and checked manually.

3.5. SXRD data reduction: advanced

Several alternative regimes were considered and tested. A brief overview is provided.

An automated background analysis is possible, in which one evaluates all of the different fitting options and selects the lowest R factor[3] accordingly. An approach for 'turning-point identification' was proposed for background identification for thin-film samples, whereby we find the first turning point (minimum between two Laue oscillations) before and after a Bragg peak and apply 'linear + Gaussian' in between these points and 'constant' elsewhere. Identification of the turning point requires one to iterate through the data set once before proceeding and is therefore slower. In addition, for thicker samples background is present past the first turning point. An alternative was to define the turning point by the Laue fringe separation. As long as the film thickness is approximately known one can estimate before the analysis which points should receive which background calculation. In principle, although these methods work, the trade-off between computational time and the number of additional data points treated prevents their implementation for 'live' structure factors.

A second expanded background was implemented to analyse the validity of the fitted background, which when placed far from the signal can be used to identify the presence of powder rings. Powder rings are identified by comparing background fits in the two bROIs. Over the size of the signal, a powder ring is approximated by a Gaussian profile multiplied by a linear profile. If a two-dimensional fit with a Gaussian profile in one dimension and a linear profile in the other was observed in both bROIs, a powder ring is probable and thus the data point was marked 'd'. Powder rings also provide a secondary check for locating signal that leaks out of the sROI into the bROI by sampling regions far from any expected signal. If the two fits are not consistent it is probable a signal leak has occurred, and thus consistency between a background fit to these regions and the bROI confirms if the signal identified is suitable.

The idea of an overall sROI was tested. All good data point sROIs were added together and used as the estimated sROI should one of the criteria fail. An alternative was a moving average approach: the last five ROIs are kept, averaged and used as the next sROI. This was both successful and destructive: in some instances we could go on to recover points previously unassigned, and in other instances, previously assigned data points were now erroneous.

How do we identify the powder rings? Scattering from individual crystallites in powder rings was avoided by enforcing consistent pixels between consecutive signals, such as the central pixel on the detector, as discussed previously.

How can we identify multiple clusters/split signals? A split signal, *e.g.* due to sample miscut, is characterized by multiple footprints of the same shape. It is common to work with miscuts of $<0.1^\circ$ and thus split signals are not encountered regularly. However, for the general application of the method one could envisage identifying matching signal shapes *via* a convolution-based cross correlation.[4]

[3] The reduced χ^2 function used to fit the background is detailed by Schlepütz (2009).

[4] Given two images A and B, of dimensions $n \times n$, a convolution-based cross correlation is achieved using a fast Fourier transformation (FFT), whereby the location of the maximum of $\mathrm{FFT}^{-1}[\mathrm{FFT(A)} \times \mathrm{FFT(B)}]$ defines the required shift to best correlate the images (Press *et al.*, 1988).

A lot of the constraints described have an effect on each other, and it is thus difficult to separate the constraints into truly objective steps. Hence they must be tailored to individual samples. Inevitably there will always be a human check required at the end of each data set to reject or confirm the 'd'ubious points. More complex methods become useful when data sets approach tens of thousands of structure factors.

4. Results

4.1. Automated extraction of SXRD

The differences between an automated and manual extraction are shown in Fig. 6. The data were measured in a beryllium dome, under modest vacuum, on a sample consisting of an $LaNiO_3$ thin film grown on an $SrTiO_3$ substrate. The majority of the data points lie within 5% of each other; the largest deviations occur near the turning points when the signal is weak and the associated errors are large. The total difference between the data sets is ~6%. This is significant but within typical error bars for data sets of this type. A visual inspection of the automated results confirms they are reasonable. A manual extraction requires approximately 30 min, an automated extraction less than 2 min.

All data points where the error is larger than 10% are identified as dubious (~10%) or bad (~4%) by the automated extraction so must be dealt with manually. Thus 86% of the data set has been dealt with successfully in this case. The number of dubious data points drops to 5% with the use of a Kapton dome sample environment and consequent removal of the diffuse background and powder rings produced by the polycrystalline beryllium.

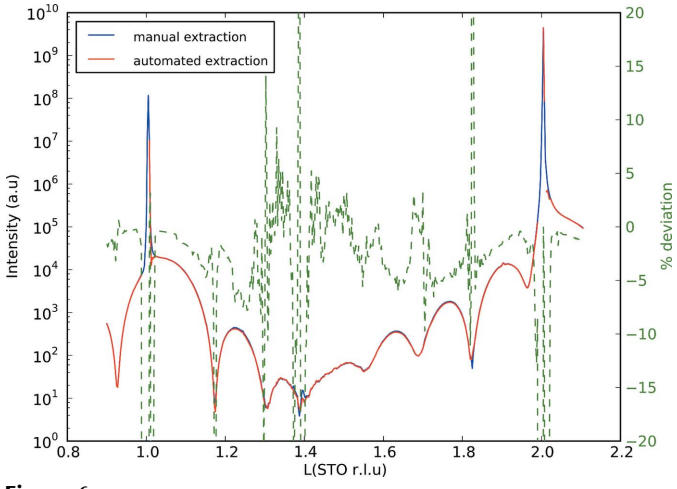

Figure 6
The corrected intensity (arbitrary units) extracted manually (blue) and automatically (red) and their relative deviation for a (00*L*) CTR from an $LaNiO_3$ thin film grown on an $SrTiO_3$ substrate (STO), measured under vacuum in a beryllium dome.

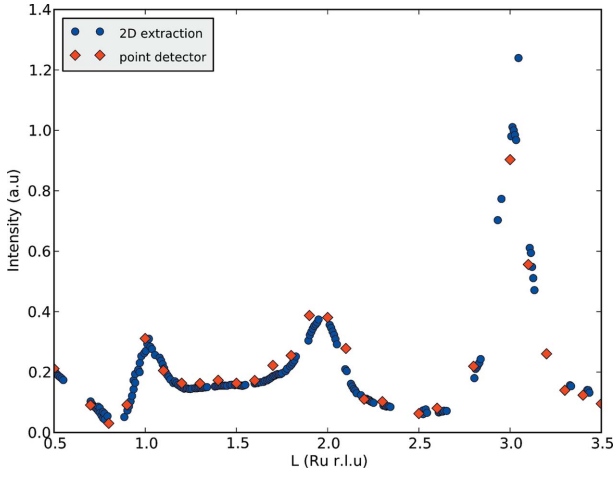

Figure 7
An example of data extracted from a rocking curve scan using an area detector as a point detector, compared with the same data but using the area detector, for the superstructure signal from graphene on ruthenium (Martoccia *et al.*, 2010).

4.2. Two-dimensional rocking curves

To provide an example (see Fig. 7) of the gains to be made, we applied this method to 27 rocking curves taken at 0.1*l* steps along a superstructure truncation rod from a graphene on ruthenium(0001) sample (Martoccia *et al.*, 2010). Not only is the number of structure factors obtained increased by a factor of ten, but the reliability is tested, as the background correction is more robust and outliers more easily identified.

A point to stress is that often those familiar with traditional SXRD methods use area detectors as point detectors, insofar that only the signal at the detector centre is extracted, while all the remaining information in the detected image about neighbouring regions of *k* space is thrown away (Mariager *et al.*, 2009; Schlepütz *et al.*, 2011). We expect this method to improve efficiency by over an order of magnitude, particularly for weak and diffuse signal, such as superstructure signal due to octahedral rotations in ultra-thin films (May *et al.*, 2010). The method also allows one to rapidly obtain reliable structure factors at low *l*, where the open slit geometry breaks down.

5. Conclusion

We have demonstrated an efficient cluster-algorithm-based approach to signal identification in two-dimensional diffraction data. The cluster method can be easily extended to three-dimensional distributions, *i.e.* 6 and 26 voxels for the respective neighbourhood criteria. The algorithm is thus applicable to any image processing field, for example X-ray free electron lasers, which generate very large volumes of data, often with many thousands of signals to analyse in an individual image (Pedrini, 2012).

The cluster method was applied in the context of SXRD to automate the extraction of structure factors, thus reducing significantly this bottleneck towards a structural solution, minimizing the impact of human error, improving the efficiency of the technique by orders of magnitude and hence making the technique more accessible. It should be noted that the method, although not definitive, is sufficiently flexible to

apply to many different systems, but at some level human input is always required. The automated approach described here is a sound objective starting point.

An experimental method was outlined to improve measurements made in low-l geometries, improving the scope of the technique and demonstrating the often overlooked information gained from two-dimensional detectors.

Support of this work by the Schweizerischer Nationalfonds zur Förderung der wissenschaftlichen Forschung and the staff of the Swiss Light Source is gratefully acknowledged. This work was performed at the Swiss Light Source, Paul Scherrer Institut. We would like to thank C. M Schlepütz for development of the *Scananalysis* GUI.

References

Bjoerck, M., Schlepütz, C. M., Pauli, S. A., Martoccia, D., Herger, R. & Willmott, P. R. (2008). *J. Phys. Condens. Matter*, **20**, 445006.
Bolotovsky, R., White, M. A., Darovsky, A. & Coppens, P. (1995). *J. Appl. Cryst.* **28**, 86–95.
Brönnimann, C., Eikenberry, E. F., Horisberger, R., Hülsen, G., Schmitt, B., Schulze-Briese, C. & Tomizaki, T. (2003). *Nucl. Instrum. Methods Phys. Res. Sect. A*, **510**, 24–28.
Filhol, A., Thomas, M., Greenwood, G. & Barthelemy, A. (1983). *Position-Sensitive Detection of Thermal Neutrons*, edited by P. Convert & J. B. Forsyth. London: Academic Press.
Ford, G. C. (1974). *J. Appl. Cryst.* **7**, 555–564.
Kabsch, W. (1988). *J. Appl. Cryst.* **21**, 916–924.
Kossel, W., Loeck, V. & Voges, H. (1935). *Z. Phys.* **94**, 139–144.
Lehmann, M. S. & Larsen, F. K. (1974). *Acta Cryst.* A**30**, 580–584.
Livet, F., Bley, F., Mainville, J., Caudron, R., Mochrie, S., Geissler, E., Dolino, G., Abernathy, D., Grübel, G. & Sutton, M. (2000). *Nucl. Instrum. Methods Phys. Res. Sect. A*, **451**, 596–609.
Mariager, S. O., Lauridsen, S. L., Dohn, A., Bovet, N., Sørensen, C. B., Schlepütz, C. M., Willmott, P. R. & Feidenhans'l, R. (2009). *J. Appl. Cryst.* **42**, 369–375.
Martoccia, D., Björck, M., Schlepütz, C. M., Brugger, T., Pauli, S. A., Patterson, B. D., Greber, T. & Willmott, P. R. (2010). *New J. Phys.* **12**, 043028.
MathWorks (2012). MATLAB. Version 7.14.0.739 (R2012*a*). The MathWorks Inc., Natick, MA, USA, http://www.mathworks.ch/products/matlab/.
May, S., Kim, J.-W., Rondinelli, J., Karapetrova, E., Spaldin, N., Bhattacharya, A. & Ryan, P. (2010). *Phys. Rev. B*, **82**, 014110.
Munkholm, A. & Brennan, S. (1999). *J. Appl. Cryst.* **32**, 143–153.
Pauli, S. A., Leake, S. J., Björck, M. & Willmott, P. R. (2012). *J. Phys. Condens. Matter*, **24**, 305002.
Pedrini, B. (2012). Personal communication.
Press, W. H., Flannery, B. P., Teukolsky, S. A. & Vetterling, W. T. (1988). *Numerical Recipes in C: The Art of Scientific Computing.* Cambridge Univerity Press.
Rivers, M. (2007). *Trajectory Scanning with the Newport MM4005 and XPS Motor Controllers*, http://cars9.uchicago.edu/software/epics/trajectoryScan.html.
Rossmann, M. G. (1979). *J. Appl. Cryst.* **12**, 225–238.
Roth, M. & Lewit-Bentley, A. (1982). *Acta Cryst.* A**38**, 670–679.
Saldin, D., Harder, R., Shneerson, V. & Moritz, W. (2001). *J. Phys. Condens. Matter*, **13**, 10689–10707.
Schlepütz, C. M. (2009). PhD thesis, pp. 1–281, University of Zurich, Switzerland.
Schlepütz, C. M., Herger, R., Willmott, P. R., Patterson, B. D., Bunk, O., Brönnimann, Ch., Henrich, B., Hülsen, G. & Eikenberry, E. F. (2005). *Acta Cryst.* A**61**, 418–425.
Schlepütz, C. M., Mariager, S. O., Pauli, S. A., Feidenhans'l, R. & Willmott, P. R. (2011). *J. Appl. Cryst.* **44**, 73–83.
Schoenborn, B. P. (1983). *Acta Cryst.* A**39**, 315–321.
Sjölin, L. & Wlodawer, A. (1981). *Acta Cryst.* A**37**, 594–604.
Spencer, S. A. & Kossiakoff, A. A. (1980). *J. Appl. Cryst.* **13**, 563–571.
Team, T. M. (2010). *Medipix*, http://medipix.web.cern.ch/medipix/pages/medipix3.php.
Vlieg, E. (1997). *J. Appl. Cryst.* **30**, 532–543.
Wilkinson, C., Khamis, H. W., Stansfield, R. F. D. & McIntyre, G. J. (1988). *J. Appl. Cryst.* **21**, 471–478.
Wlodawer, A. & Sjölin, L. (1982). *Nucl. Instrum. Methods Phys. Res.* **201**, 117–122.
Yacoby, Y., Sowwan, M., Stern, E., Cross, J., Brewe, D., Pindak, R., Pitney, J., Dufresne, E. B. & Clarke, R. (2003). *Physica B*, **336**, 39–45.