# Finding Single Crystal Peaks

Samuel Jackson, Michael Hart, and Anders Markvardsen

March 16, 2017

## 1 Problem Statement

The WISH instrument runs approximately 60 single crystal diffraction experiments a year. Identifying and correctly indexing single crystal peaks in the diffraction pattern is the first step in the analysis of a sample under study. Identifying the location of both strong and weak crystal peaks in the diffraction pattern allows the experimentalist to pin down the orientation of the sample. The location, shape, and integration values of peaks is also central to deducing structural information in subsequent analyses.

Unfortunately, finding peaks in a noisy background is a difficult problem. A significant proportion of instrument scientist and user time is spent adjusting the output of the 3D peak finding algorithm in Mantid. The existing approach focusses on identifying high density regions in the 3D Q (or if indexed, HKL) space of a particular run and uses a simple intensity & density thresholding approach as the acceptance criteria for peak selection.

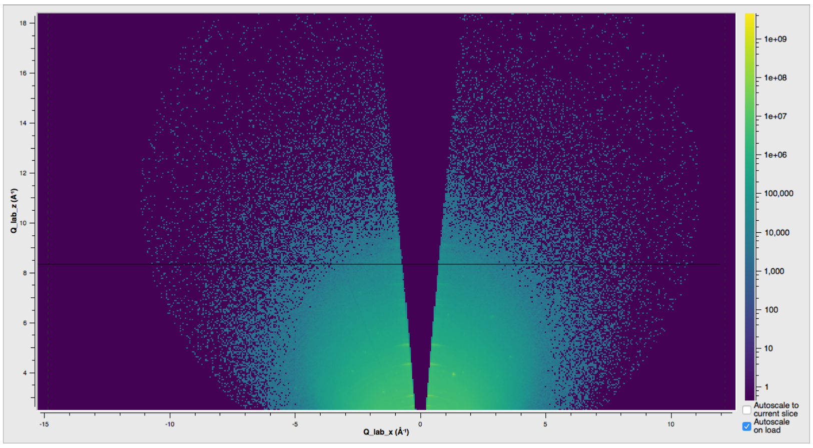
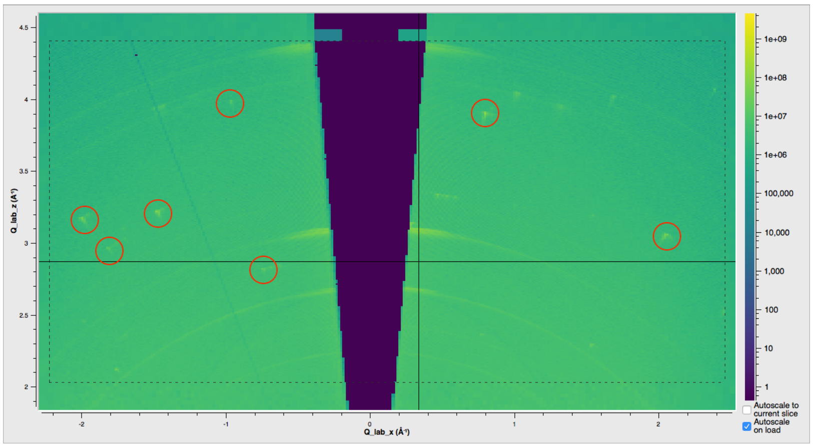


Figure 1: A slice through the $Q_y$ axis of WISH Q space



□ Autoscale to current slice

☑ Autoscale on load

Figure 2: Single crystal diffraction Peaks in WISH Q space originating from the sample. Also shown are powder diffraction rings which are typically generated by the sample container.

This approach frequently fails to produce adequate output for several reasons:

- False positive peaks. This is usually due to a combination of complicated background and artifacts in the diffraction pattern (such as powder diffraction rings).
- False negative peaks. Usually weak peaks which are not detected by the algorithm due to noise or a threshold set too high to counter the background.
- Double peaks, where two instances of the same peak are seen as two separate peaks.

Despite these complications, a human expert can easily visually identify peaks in the pattern, providing motivation that machine learning techniques may have more success where the existing approach fails.

An early task in this project will likely be to examine how, if possible, we can model the noise generated from the sample environment. Reliably being able to model this noise could be an advantagous preprocessing step that drastically simplify the problem of peak identification and will almost certainly be needed for the generation of synthetic datasets ( see section 3).

## 2 Automated Solutions

Ideally the process of identifying single crystal peaks should be as automated as possible, allowing scientists to focus on the more interesting analysis of their experiment. An ideal peak finder would likely have the following properties:

- A low false positive & false negative rate for both strong and weak peaks.
- A confidence estimate in the prediction. How certain is the algorithm that this particular peak is a peak?

- Information relating to the shape & size of identified peak. Including perhaps classification.
- Continuous learning. The model should be able to take in corrections to predicted results and use this to improve successful detection.

The following sections outline two approaches that could be taken as starting points to develop a machine learning solution to the problem in the previous section. Both of these starting points are deliberately left loosely defined. These are only suggestions of where to begin and will likely be adapted and experimented with as the project progresses.

## 2.1 Hand crafted features Approach

One approach would be to use hand crafted features based on tried and tested computer vision methods. Such an approach would most likely identify regions of interest (ROIs) based on (for example) blob detection methods. An ROI detection method could be build ontop of the existing peak detection method, but use the located points as the centre of an ROI. Discriminating features based on these ROIs, such as shape, intensity, or texture based metrics, can then be used to train a binary classification model. This could be any one of the commonly used models in machine learning such as a SVM or Random Forest. In fact it would be best to train a number of models on the same features and compare their performance.

Classifiers such as these can easily be modified to give a confidence estimate on the predicted class. New images could then be shown to the trained system, the ROIs & features calculated, and the ROIs with similar metrics would be predicted as containing a peak (or not).

An advantage of using this as a starting point is that such a system would likely be much easier to develop and train. However, the performance of such a system will be highly dependant on the method used to select ROIs and the choice of features computed from those ROIs. Choosing this method would mean initially spending time investigating which methods will best suit the problem data.

## 2.2 Neural Network based Approach

Another approach would be to train a neural network directly on the data either on the entire volume itself or on 2D slices (see section 3). The architecture of the neural network would have to be explored experimentally, but the work in refs. [10] and [7] on classifying x-ray data could provide a potential a starting point.

An important point to consider when developing a neural network based approach is that this problem requires both object identification (“there are peaks”) and object detection (“the peaks are here”). Successful detection of where an object is in an image is a computationally intensive problem, but recent work in the neural network community has produced some effcient solutions to mutliclass object detection problems based on regression networks [9] and region proposal networks [7], [6], [8].

One potential solution would be to have two neural networks in a pipeline. One network trained to propose regions of interest, either outputting coordinates of a bounding box or outputing scores for individual pixels. A second network or even another type of classifier could then be used with the generated ROIs as input. The output of this second network would then be a classification of where the ROI does or does not contain a peak.

## 3 Datasets & Representation

The current method for peak finding searches across 3D bins in the Q space of a particular run. Many domains which use the methods described in the preceeding section are applied directly to 2D images. For our problem we could either choose to develop a system which works directly on the 3D volume or slice the dataset along a given axis to form a series of 2D images. The latter has the advantage of producing a large dataset set from only a few sample runs, with the caveat that some peak examples in adjacent slices would not necessarily be independent of one another. Ideally both methods should be investigated if time and resources allow.

The other consideration with the representation is which space to use for training and testing the model? Both will need to be the same for an apples to apples comparison. Q space seems the most logical here. Without necessarily knowing the orientation (UB) matrix in advance one cannot transform the data to HKL space. Using detector space would negate the problem of first having to convert the data to Q (which is computationally slow) but has the disadvantage of being instrument specific and more difficult to relate adjacent spectra.

The lack of labelled data currently available is also an issue. Scientists typically perform their peak identification on the fly during or after and experiment. There is currently no single, consistently labelled ground truth dataset which can be used. However, WISH scientists note that they do have some indexed peaks stored which could possibly be used, but the extent and detail of the data is currenty unknown. This lack of data is a major sticking point for a machine learning approach. A good volume of diverse, labelled diffraction data is necessary for supervised learning methods. This is where being able to continuously learn from user input is likely to be useful. Existing labelled data could be supplemented by transforming it in some way to produce “new” examples. For example, taking a ROI with a single peak in it and rotating it by 90 degrees.

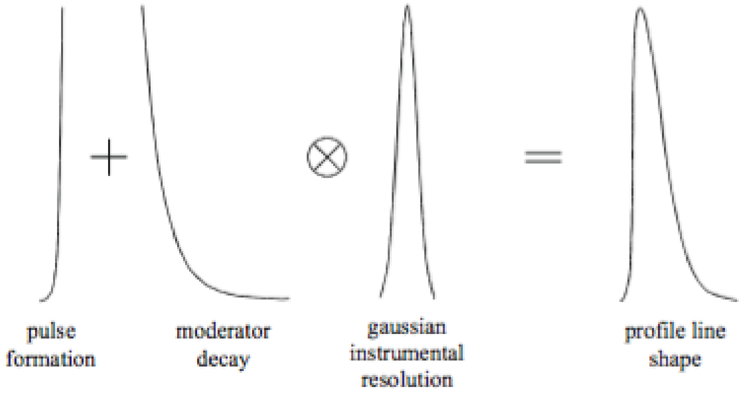


Figure 3: Components making up the peak line shape in time-of-flight diffraction.

One way around this problem could be to use artificial or simulated data. Unlike many image analysis problems, it could be quite easy to simulate training examples. Single crystal peaks at the most fundamental level be defined by the instrument resolution function and functions for pulse formation and moderator decay (see figure 3). It should be relatively simple to generate data that closely resembles real experimental data using a combination of know crystal lattices, instrument resolution, and a UB matrix. Another option is to use of software such as McStas [2] or MCViNE [3] could be used as starting points for generating datasets. How accurately simulated data would match features of real data has yet to be determined. The key question here is that simulated data must be accurate enough to real data and also as varied as the real data in order to avoid overfitting to the simulated data.

Another issue is that both the exisiting technique and the proposed versions above are limited as they can only incorporate visual data. More advanced methods which include heuristics based on, for example, the sample envrioment used could also possibly be incorporated. Modelling of this may be helpful to reduce noise in the data, and this could easily be explored at the same time as attempting to accuratly simulate TOF data.

## 4 Implementation Considerations

For initial development of the system the programming language Python would be strongly recommended. This is for a number of reasons:

- Mantid, used for reducing the raw data, exposes a Python API.
- Python is widely used in the scientific and machine learning communities.
- Python has a wide variety of tried and tested image processing [4] and machine learning [5] [1] libraries.
- Python easily interoperates well with C/C++ meaning custom high performance functionality can easily be included.
- Authors have familiarity with the language.

Additionally, there are some hardware considerations that must be addressed. WISH data is typically fairly large ( 800Mb on disk, larger in memory) and the transformation to Q space is computation intensive. This means that the any development machine need sufficient memory and processing power to handle these constraints. Furthermore, machine learning models, and in particular neural network models are typically also computationally expensive to train often requiring the use of GPUs for realistic training times. When larger computational resources are required, one option would be to use the HPC resources maintained by STFC.

## 5 Related Work

We are unaware of any directly comparable work to the problem described here. The closest existing work found was that of refs.[7] and [10] on the classification of x-ray scattering patterns using deep neural networks. While their work has some superficial similarities to this problem, it only classifies the pattern into a discrete category and says little about localisation of interesting artifacts used in the classification.

## 6 Conclusion

In conclusion, the problem of finding peaks in single crystal diffraction data has been outlined and the difficulties of the existing system have been listed. Two feasible alternative approaches based on the machine learning paradigm have been presented. Based on the specification of the task at hand and its intersection with similar challenges in the image processing & machine learning communities we feel that such a system would be feasible to build and could provide superior performance to the existing method.

However, it is worth emphasising that building such a system would not be without challenges. The major hurdles which we feel must be overcome are obtaining/creating enough labelled data for training and the computational expense of preprocessing WISH data to Q space during the training stage. From there we feel that a diverse number of approaches should be considered and trialed to get best understanding of methods which work best with the data at hand.

Based on an initial analysis of this problem we feel that the next steps towards building an automated peak finding system in order of importance would be:

- Creation of ground truth dataset with a high variety of peak strengths & shapes from a multitude of representative samples. As mentioned previously the lack of good labelled data is a significant problem. We feel that due to the well defined and simplistic nature of peaks our best option is to turn to simulating the data. Providing this is accurate to the real data this would give us plenty of training examples to work with and also allow us to control the size of the data to make prototyping quicker & easier.

- Run the current peak finding method both on real experimental data and on the simulated dataset and evalute the true & false positive/negative rates achieved. This gives us a known baseline which we can use to measure the performance of subsequent prototypes against.

- Begin to develop simple prototypes of the system using as much “out of the box” software as possible. Trying many different types of technqiues will allow us to compare and contrast each approach and hopefully allow us to identify the most feasible approaches as quickly as posible. Using existing software will aid how quickly we can test candidate solutions while adding reassurance of the correctness of the implementations used.

While it is easy to outline the next steps towards implementation, it is much more difficult to esimate the amount of time & effort required, particularly for the third point above. To a certain extent decisions need to be made after experimenting with the data and improving any models based on results and performance. All we can say is that the three points above are perhaps the bare minimum steps required to get a simple prototype potentially capable of out performing the exisitng method up and running.

This is to the best of our knowledge the first attempt at ISIS to use machine learning for improving data processing. We feel that machine learning has a lot to offer ISIS, and the success of this project will likely be the seed for further utilisation of such techniques at ISIS, potentially greatly improving the quality and speed by which data can be processed.

## References

[1] Keras website. `https://keras.io/`. Accessed: 2017-01-30.

[2] McStas website. `http://mcstas.org/`. Accessed: 2017-01-30.

[3] MCViNE website. `http://web.archive.org/web/20080207010024/http://www.808multimedia.com/winnt/kernel.htm`. Accessed: 2017-01-30.

[4] Scikit-Image website. `http://scikit-image.org/docs/dev/`. Accessed: 2017-01-30.

[5] Scikit-Learn website. `http://scikit-learn.org/stable/`. Accessed: 2017-01-30.

[6] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

[7] M Hadi Kiapour, Kevin Yager, Alexander C Berg, and Tamara L Berg. Materials discovery: Fine-grained classification of x-ray scattering images. In *Applications of Computer Vision (WACV), 2014 IEEE Winter Conference on*, pages 933–940. IEEE, 2014.

[8] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2015.

[9] Pierre Sermanet, David Eigen, Xiang Zhang, Michaël Mathieu, Rob Fergus, and Yann LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. *arXiv preprint arXiv:1312.6229*, 2013.

[10] Boyu Wang, Kevin Yager, Dantong Yu, and Minh Hoai. X-ray scattering image classification using deep learning. *arXiv preprint arXiv:1611.03313*, 2016.