Localization of Signal Peaks in Photon Science Imaging

Daniel Becker
University of Applied Science
Berlin, Germany
Email: beckerd@htw-berlin.de

Achim Streit
Steinbuch Center for Computing
Karlsruhe Institute of Technology, Germany
Email: achim.streit@kit.edu

***Abstract*—In order to get insights into the atomic structure of tiny samples like proteins or viruses, X–Ray microscopy can be utilized. Here, an X–Ray laser illuminates crystallized samples and the scattered light is captured by a detector device in the form of an image. Due to limitations of the experimental setups, these images not only contain the relevant data, but are contaminated by noise from various sources. This makes it hard to isolate useful images containing enough valid signals to be used for further research. In this article, we explore the different sources of noise and propose a series of techniques for identifying valid signals in photon science. The outcome of our algorithm is then compared to predictions obtained by the tools currently in use.**

***Keywords*—image processing; photon science; big data; signal processing; X–ray microscopy; nanocrystallography.**

I. Introduction

In photon science, tiny crystallized samples are illuminated by X–ray laser light pulses in order to explore their internal structure. The intensity distribution of the diffracted light is captured by a detector device which takes an image for every laser pulse. The images may show bright peaks ("Bragg spots") whose locations are essential for unfolding the internal structure of the probe.

Since X–rays are used, the samples are destroyed withing femtoseconds due to the intensity of the laser. Consequently, many samples have to be analyzed for the unfolding analysis. In order to efficiently do this, experiments are constructed with a probe transportation system in place. Fig. 1 shows the setup of the LCLS experiment at Stanford[1]. Here, a jet stream is used to transport the probes across the light source. This makes it possible to move a high amount of probes, which is necessary for the 120 Hz repetition rate of the laser. This results in 120 taken images per second. However, since it is not possible to synchronize the stream of probes with the laser pulses, only an order of 5% of the samples are illumintead in a way that allows further analysis. Consequently, up to 95% of the images are useless for further research[2].

The resolution of the detector device is 2.3 MP at 14-bit depth. This results an image size of about 4 Megabytes. At a rate of 120 Hz, this results in a data volume of about 1.8 TB/h. Currently, all data is stored offline for later analysis. This is feasable due to the rather low amount of data.

But this won't be an option in experiments currently being built. For example, the European X–ray Free Electron Laser (XFEL) will operate at an image repetition rate of 27,000 Hz [3]. Here, even if the detector is kept at the same resolution, the amount of data will increase by factor 225. This leads to the problem, that storing data offline is no longer an option. To achieve the full potential of this experiment, new solutions have to be explored in order to reduce the amount of data actually stored for later analysis.

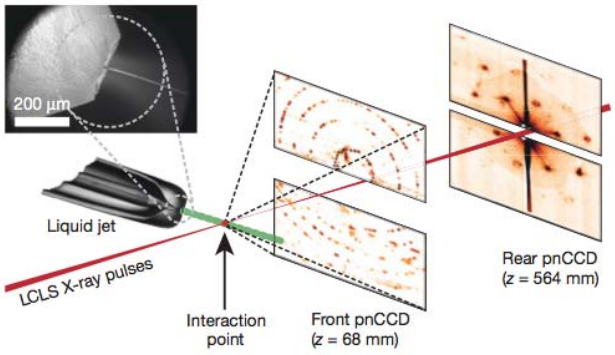


Fig. 1. Nanocrystallography. X-ray pulses from a free-electron laser (FEL) interact with nanocrystals flowing in a buffer solution. The detector records the resulting diffraction patterns [7].

In [4] we introduced a simplistic and fast neural network for weeding out entirely blank images during the data collection phase, being as close to the detector as possible. The problem remains to handle images containing some data, but which are still not usable for more sophisticated analysis.

In this article, the next step in the analysis chain is considered, namely determining the position of Bragg spots with a view to further pursuing data reduction in near–real time. Bragg spot identification is made difficult due to noise smeared non–uniformly over the image due to the diffraction of laser light by the buffer solution. In addition, there is a noise contribution due to stuck or broken pixels in the detector device. In this article, we present three connected techniques for localizing Bragg spots. Firstly, noise removal based on a convolution technique, secondly, edge detection using the Sobel operator [5] and, thirdly, an algorithm is suggested for finding local clusters within an image. LCLS data from three different samples are used to verify our approach. The results are compared with results obtained by the Cheetah toolkit, the standard software for Bragg spot finding at LCLS [6].

A. Related Work

In the article *Crystalline object evaluation by image processing* [8] Billingsley et al. use the Sobel operator for

978-1-4799-8713-9/15 $31.00 © 2015 IEEE
DOI 10.1109/UKSim.2015.35

verifying whether the crystallization of protein samples was successful. During the growth process of the crystal, images are taken and an edge detection is performed. This information is used for finding connected lines within an image which, in turn, indicate whether a crystal will generate exploitable diffraction patterns. The decision process itself is carried out by a support vector machine (SVM).

The currently used solution for pre–selecting images from the LCLS experiment is called "Cheetah"[6]. It uses an Algorithm called "hitfinder" to analyze given images for Bragg spots. The algorithm itself only looks for a certain amount of connected pixels above a predefined intensity threshold. In addition, several optimization techniques have been incorporated to compensate for as much background noise as possible. The results of the "Cheetah" software are used as a reference to compare our proposed algorithm against.

II. Data & Methods

A. Test Data

Diffraction data from three crystallized probes are analyzed:

- the protein *Cathepsin B* (CatB) [9],
- the *5-Hydroxytryptamine receptor 2B* (5HT-2B) [10],
- the *granulovirus polyhedron* (GV) [11].

For each probe, 25 indexable and 25 non–indexable images are selected. An image is indexable if a Fourier series index can be assigned to each Bragg spot of a diffraction pattern. Bragg spots and the electron density of a sample are related (in lowest order) by a Fourier transformation, see e.g. [12].

B. Geometry

The detector device is composed of multiple panels (see Fig. 2). Those individual panels are organized into four quadrants. These quadrants are able to slide in and out relative to the center of the detector. This is nesessary to compensate for different distribution angels of the scattered light by the probes, since this can vary due to the transportation liquid used as well as the structure of the probe researched.

However, this results in varying coordinates for each detector pixel within an image. To work around this, a separate geometry file is created for each experiment, containing the physical coordinates for each pixel for this experiment. The images themselves just contain the sub–images of the panels organized in a grid to which the geometry file can be applied to retrieve the physically correct image. Since this is not necessary for our proposed algorithms it is not applied within our experiments. The same is true for the Cheetah toolkit, to which we compare our results. Nonetheless, it is applied to the images in this article to support the illustration of our algorithms.

C. Data Normalization

The readout of the pixels is corrected for unphysical data. The detector collects 14-bit unsigned integer data per pixel [13]. Negative or very large ($>$ 16,000) readout numbers indicate broken or stuck pixels and are therefore set to zero.

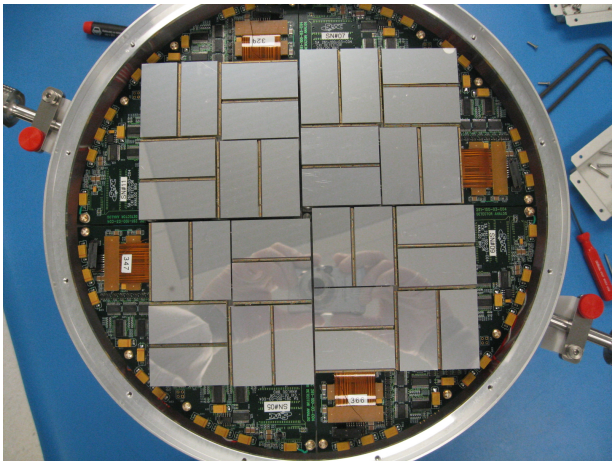


Fig. 2. The CSPad detector device used at the LCLS experiment at Stanford.

Let $I_i^{(\mathrm{raw})}$ be the readout number of the i–th pixel of an analyzed image. The quantity

$$I_i = I_i^{(\mathrm{raw})}\,\theta(I_i^{(\mathrm{raw})})\,\theta(16{,}000 - I_i^{(\mathrm{raw})})$$

represents the *intensity* of the i–th pixel. $\theta(p)$ is the unit step function (zero for negative p, one for non–negative p). The intensity is normalized to the interval, $0 \leq I_i \leq 16{,}000$.

Each detector pixel i can also be identified by two integers (x, y) characterizing the position of a pixel in the detector plane where $0 \leq x, y < \dim(I)$.

D. Average Subtraction

Laser light is always diffracted when propagating through the buffer solution resulting in an overall background. This background can be removed to a large extend from an image by analyzing a series of known blank images and determining an average noise for each pixel of the image.

To estimate the average background noise, N=500 blank images are analyzed in advance and for every pixel i the *mean noise*

$$I_i^{(\mathrm{noise})} = \frac{1}{N}\sum_{n=1}^{N} I_{n,i}^{(\mathrm{blank})} \tag{1}$$

is determined, where $I_{n,i}^{(\mathrm{blank})}$ denotes the intensity of the i–th pixel in the n–th blank image (corrected for unphysical readout data, see Sec. II-C).

The influence of background effects is explored by considering the corrected intensity

$$\widetilde{I}_i = (I_i - I_i^{(\mathrm{noise})})\theta(I_i - I_i^{(\mathrm{noise})}). \tag{2}$$

It should be noted that the noise subtraction depends on the kind of fluid used to transport the crystals. In this article, it is assumed that the background is static during a measurement period.

E. Noise Reduction

Diffraction images contain noise from different sources. In this section we concentrate on sharply localized noise affecting only single detector pixels. In contrast, Bragg spots are usually distributed over several pixels.

The probability for misidentifying a single spot pixel as a Bragg spot can be reduced considerably by distributing its intensity over the neighboring pixels by using convolution techniques, well known in image processing [14]. Let us introduce the three–dimensional matrix

$$K = \frac{1}{9}\begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}.$$

By convoluting K with an image I

$$I^{(\text{noise\_reduced})} = I * K$$

where the convolution $(*)$ is defined by

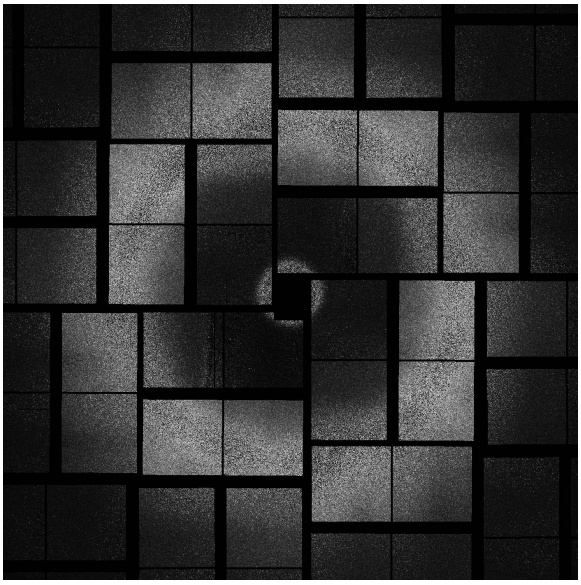
$$I^{(\text{noise\_reduced})}_{x,y} = \sum_{m=1}^{\dim K} \sum_{n=1}^{\dim(K)} I_{x-m+2,y-n+2} K_{m,n},$$

the strength of a single pixel noise at the position (x, y) can be reduced.

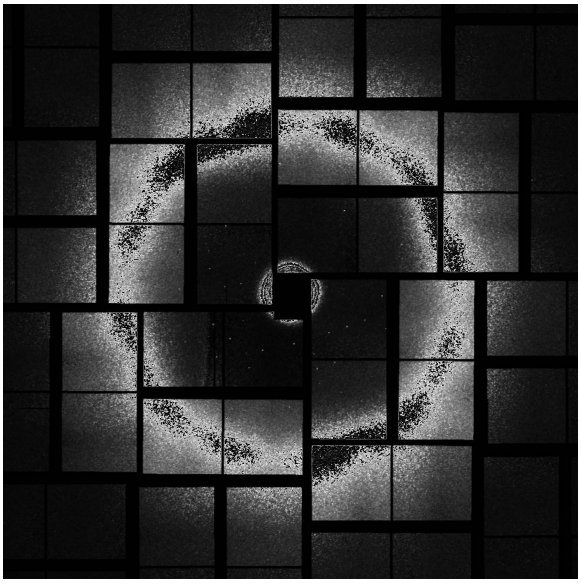
For consistency, the matrix I is set to zero beyond its boundaries, i.e. $I_{x,y} = 0$ for $x, y < 0$ and $x, y \geq \text{size}(I)$.

Note that there is no overall intensity loss due to the convolution operation as the normalization of the matrix K is chosen such that the sum of its matrix components is one. After convolution, the intensity of isolated single–pixel spots is damped considerably and approximately comparable with the average noise in the neighborhood.

An example of the usage of convolution can be seen in Fig. 3. It can be seen that single pixel noise is drastically reduced and as a result, spots composed of more than one pixel are now standing out against the background. This is best visible in the area around the center of the image.



(a) before



(b) after

Fig. 3. Example of an image before (3a) and after (3b) convolution with the matrix K.

F. Edge Detection

Bragg spots differ in size, shape, and intensity. However, what many spots seem to have in common is the shape of their intensity distribution. This characteristic feature can be harnessed for identifying Bragg spots by analyzing their boundaries. A standard technique for detecting edges within an image is using the horizontal and vertical Sobel operators [15]

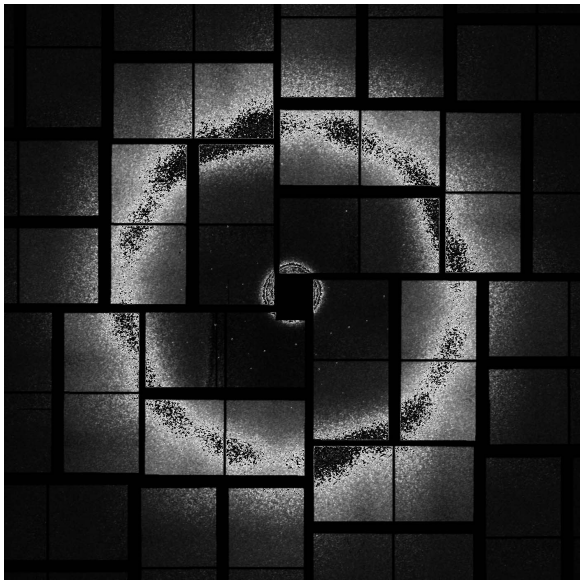
$$S_h = \begin{pmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{pmatrix}, \quad S_v = \begin{pmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{pmatrix}.$$

They are convoluted with an image to determine horizontal and vertical intensity changes. The quantity

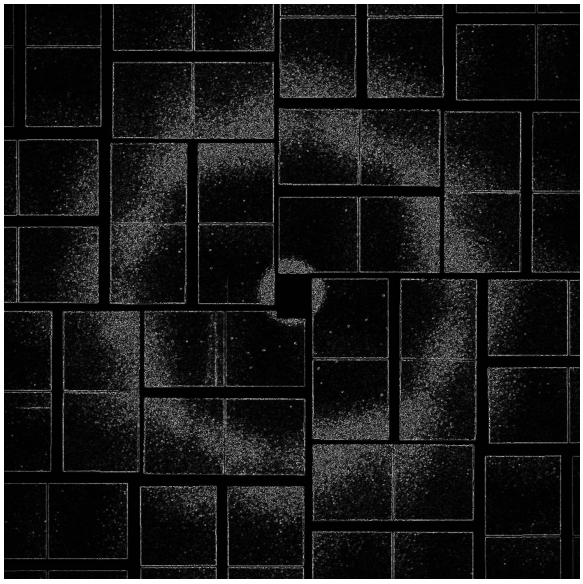
$$\Delta I_{x,y} = \sqrt{\left(I^{(\text{noise\_reduced})} * S_h\right)^2_{\text{x,y}} + \left(I^{(\text{noise\_reduced})} * S_v\right)^2_{\text{x,y}}}$$

can be considered as a measure for the strength of the local intensity change at the pixel position (x, y) of the image I. Large values for $\Delta I_{x,y}$ indicate an edge at (x, y).

Fig. 4 shows the results of the application of edge detection to a noise–reduced image. It can be seen that the spots around the center now stand out even more clearly. Furthermore, the previously still noisy water halo has been dampened as well, making the spots stand out clearer in comparison.



(a) before



(b) after

Fig. 4. Example of an image before (4a) and after (4b) application of edge detection.

G. Cluster Detection

After applying an edge detection, Bragg peaks can be viewed as a cluster of connected pixels. This means, a simple cluster finding algorithm can be used to find connected pixels within the image, see Fig. 5.

First, an empty matrix is created using the dimensions of the analyzed image (cluster matrix). Then, each pixel of the original image is examined. For each of these pixels, the algorithm looks at each surrounding pixel within a defined distance. If the intensity of one of these adjacent pixels is higher than a defined threshold, the intensity of the corresponding original pixel in the cluster matrix is increased.

```
1: procedure CLUSTERDETECTION(image)
2:   clusterMatrix ← dim(image)
3:   maximumDistance ← 10
4:   brightnessThreshold ← 400
5:   for x = 0; x < size(image); x++ do
6:     for y = 0; y < size(image); y++ do
7:       for i = x - maximumDistance; i + maximumDistance; i++ do
8:         for j = x - maximumDistance; j + maximumDistance; j++ do
9:           if image[x][y] >brightnessThreshold
                 && distance(image[x][y],clusterMatrix[x][y]) then
10:            clusterMatrix[x][y] ← clusterMatrix[x][y] + 1
11:          end if
12:        end for
13:      end for
14:    end for
15:  end for
16: end procedure
```

Fig. 5. Cluster detection in pseudo–code

The cluster detection algorithm is not designed for rating the size and shape of the Bragg spots found. Since the lack of insights from different areas, e.g. from theoretical models, it is not possible to make an informed decision, whether a found cluster is too small or large to represent an actual Bragg spot. Therefore, only clusters which are obviously too big to represent a Bragg spot are removed. The threshold has been set to 1000 pixels.

Once there is a better awareness about the actual boundaries for spot sizes, the cluster algorithm can be extended so that spot candidates violating the limits are removed.

III. Results

The proposed techniques in this paper have been applied to three samples, see Sec. II-A. The Bragg spots identified by using our "cluster finder" algorithm have been compared to the results obtained by the "hitfinder" algorithm from the Cheetah toolkit.

The results of our analysis are split into four classes. One distinctive property is the removal of background photon noise before the images are analyzed. Furthermore, the analysis is applied to indexable and non-indexable images. Non–indexable means that an image is not suited for further research.

The results are shown in Tab. I and Tab. II without background subtraction and in Tab. III and Tab. IV with background subtraction applied. It can be seen that the amount of spots in common increases for the second experiment which includes background subtraction. This is due to the additional reduction of noise, which reduces the number of false positives by our algorithm.

IV. Conclusion

In this paper, we introduced techniques which, when combined, enable a detection of the majority of Bragg spots of a diffraction image. To this end, we connected the ideas of noise reduction, edge detection and cluster finding. We compared the spots found by our algorithm with the currently used Cheetah software. Depending on the noise level of the images, we are

TABLE I
FOUND SPOTS IN INDEXABLE IMAGES

	Total spots: hit finder	Total spots: cluster finder	Spots in common[1]	Spots: hit finder only[1]	Spots: cluster finder only[2]
CatB	937	1,532	89%	11%	48%
5HT-2B	1,400	2,169	68%	30%	59%
GV	1,180	1,603	42%	57%	60%

TABLE II
FOUND SPOTS IN NON-INDEXABLE IMAGES

	Total spots: hit finder	Total spots: cluster finder	Spots in common[1]	Spots: hit finder only[1]	Spots: cluster finder only[2]
CatB	0	66	0%	0%	100%
5HT-2B	827	1,768	68%	30%	59%
GV	2,015	1,287	45%	54%	53%

TABLE III
FOUND SPOTS IN INDEXABLE IMAGES WITH AVERAGE SUBTRACTION

	Total spots: hit finder	Total spots: cluster finder	Spots in common[1]	Spots: hit finder only[1]	Spots: cluster finder only[2]
CatB	937	1,427	90%	10%	42%
5HT-2B	1,400	2,169	72%	27%	58%
GV	1,180	1,093	43%	56%	19%

TABLE IV
FOUND SPOTS IN NON-INDEXABLE IMAGES WITH AVERAGE SUBTRACTION

	Total spots: hit finder	Total spots: cluster finder	Spots in common[1]	Spots: hit finder only[1]	Spots: cluster finder only[2]
CatB	0	22	0%	0%	100%
5HT-2B	827	1,768	59%	39%	68%
GV	2,015	1,204	33%	67%	14%

able to find up to 90% of the spots found by the hitfinder algorithm, as well as additional ones.

The main problem in detecting correct Bragg spots is the amount of noise introduced by various sources throughout the experiment. In this article, we explored a convolution technique as well as average subtraction which, together, are able to remove a significant part of the noise. However, it should be further investigated whether there are alternative or additional approaches for reducing the noise even more and, finally, to improve the correct recognition of Bragg spots.

The convolution based techniques for Bragg spot finding presented in this article can be parallelized, in principle, by splitting an image into several parts where a natural split is given by the subimages collected by the panels of the detector device.

To optimize the runtime of the algorithm proposed even further, it is possible to integrate multiple optimization steps into one, since the convolution operations on matrices can be applied associatively.

To remove the background noise even further, it should be researched, if it is feasible to calculate a dynamic brightness threshold for finding bright spots within an image. This could be done by evaluationg a time–series of images and adjusting the threshold dynamically.

Whether our cluster finding algorithm might be useful in identifying Bragg spots in near–real time immediately after the data taking period and, moreover, to contribute to damming the flood of data at XFEL, remains to be explored.

ACKNOWLEDGMENT

This work is supported by the Portfolio Extension of the German Helmholtz Association "Large Scale Data Management and Analysis"[16].

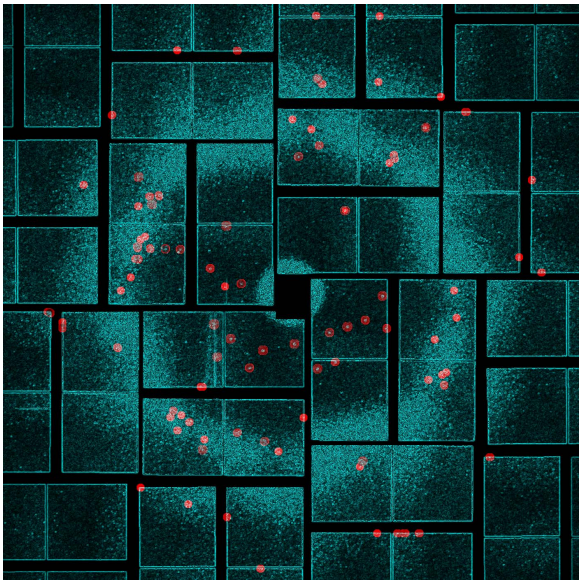
REFERENCES

[1] Y. Ding, A. Brachmann, F.-J. Decker, D. Dowell, P. Emma, J. Frisch, S. Gilevich, G. Hays, P. Hering, Z. Huang *et al.*, "Measurements and simulations of ultralow emittance and ultrashort electron beams in the linac coherent light source," *Physical review letters*, vol. 102, no. 25, p. 254801, 2009.

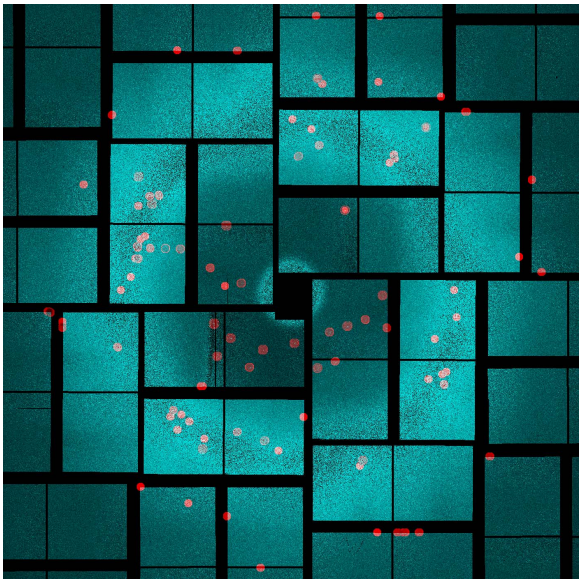
[2] S. Boutet, L. Lomb, G. J. Williams, T. R. Barends, A. Aquila, R. B. Doak, U. Weierstall, D. P. DePonte, J. Steinbrener, R. L. Shoeman *et al.*, "High-resolution protein structure determination by serial femtosecond crystallography," *Science*, vol. 337, no. 6092, pp. 362–364, 2012.

[1]Compared to hitfinder.

[2]Compared to cluster finder.



(a) Image with applied noise reduction & edge detection



(b) Image without optimization

Fig. 6. Clusters found overlaid on the optimized image (6a) and default image (6b).

[3] R. Klanner, J. Becker, E. Fretwurst, I. Pintilie, T. Pöhlsen, J. Schwandt, and J. Zhang, "Challenges for silicon pixel sensors at the european xfel," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 730, pp. 2–7, 2013.

[4] D. Becker and A. Streit, "A neural network based pre-selection of big data in photon science," in *Big Data and Cloud Computing (BdCloud), 2014 IEEE Fourth International Conference on*. IEEE, 2014, pp. 71–76.

[5] N. Kanopoulos, N. Vasanthavada, and R. L. Baker, "Design of an image edge detection filter using the sobel operator," *Solid-State Circuits, IEEE Journal of*, vol. 23, no. 2, pp. 358–367, 1988.

[6] A. Barty, R. A. Kirian, F. R. Maia, M. Hantke, C. H. Yoon, T. A. White, and H. Chapman, "Cheetah: software for high-throughput reduction and analysis of serial femtosecond x-ray diffraction data," *Journal of Applied Crystallography*, vol. 47, no. 3, pp. 1118–1131, 2014.

[7] H. N. Chapman, P. Fromme, A. Barty, T. A. White, R. A. Kirian, A. Aquila, M. S. Hunter, J. Schulz, D. P. DePonte, U. Weierstall *et al.*, "Femtosecond x-ray protein nanocrystallography," *Nature*, vol. 470, no. 7332, pp. 73–77, 2011.

[8] J. Billingsley, K. Kawabata, M. Takahashi, K. Saitoh, M. Sugahara, H. Asama, T. Mishima, and M. Miyano, "Crystalline object evaluation by image processing," *Sensor Review*, vol. 28, no. 2, pp. 143–149, 2008.

[9] L. Redecke, K. Nass, D. P. DePonte, T. A. White, D. Rehders, A. Barty, F. Stellato, M. Liang, T. R. Barends, S. Boutet *et al.*, "Natively inhibited trypanosoma brucei cathepsin b structure determined by using an x-ray laser," *Science*, vol. 339, no. 6116, pp. 227–230, 2013.

[10] J. M. Perkel, "Decoding protein structure, one femtosecond at a time," 2014.

[11] E. Chiu, F. Coulibaly, and P. Metcalf, "Insect virus polyhedra, infectious protein crystals that contain virus particles," *Current opinion in structural biology*, vol. 22, no. 2, pp. 234–240, 2012.

[12] J. Als-Nielsen and D. McMorrow, *Elements of modern X-ray physics*. John Wiley & Sons, 2011.

[13] S. Herrmann, S. Boutet, B. Duda, D. Fritz, G. Haller, P. Hart, R. Herbst, C. Kenney, H. Lemke, M. Messerschmidt *et al.*, "Cspad-140k: A versatile detector for lcls experiments," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 718, pp. 550–553, 2013.

[14] R. Keys, "Cubic convolution interpolation for digital image processing," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 29, no. 6, pp. 1153–1160, 1981.

[15] I. Sobel and G. Feldman, "A 3x3 isotropic gradient operator for image processing," *a talk at the Stanford Artificial Project in*, pp. 271–272, 1968.

[16] J. van Wezel, A. Streit, C. Jung, R. Stotzka, S. Halstenberg, F. Rigoll, A. Garcia, A. Heiss, K. Schwarz, M. Gasthuber *et al.*, "Data life cycle labs, a new concept to support data-intensive science," *arXiv preprint arXiv:1212.5596*, 2012.