JOURNAL OF SYNCHROTRON RADIATION

ISSN 1600-5775

Received 25 June 2015
Accepted 16 September 2015

Edited by M. Yabashi, RIKEN SPring-8 Center, Japan

Keywords: coherent diffraction imaging; principal component analysis; support vector machine.

Sorting algorithms for single-particle imaging experiments at X-ray free-electron lasers

S. A. Bobkov,[a] A. B. Teslyuk,[a]* R. P. Kurta,[b] O. Yu. Gorobtsov,[a,c] O. M. Yefanov,[d] V. A. Ilyin,[a,e] R. A. Senin[a] and I. A. Vartanyants[c,f]*

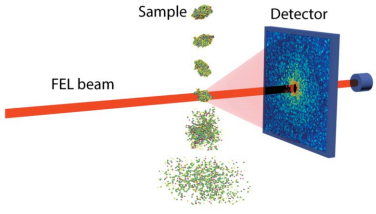
[a]National Research Centre 'Kurchatov Institute', Akademika Kurchatova pl. 1, 123182 Moscow, Russia, [b]European XFEL GmbH, Albert-Einstein-Ring 19, D-22761 Hamburg, Germany, [c]Deutsches Elektronen-Synchrotron DESY, Notkestrasse 85, D-22607 Hamburg, Germany, [d]Center for Free-Electron Laser Science, Notkestrasse 85, D-22607 Hamburg, Germany, [e]Lomonosov Moscow State University, GSP-1, Leninskie Gory, 119991 Moscow, Russia, and [f]National Research Nuclear University MEPhI (Moscow Engineering Physics Institute), Kashirskoe shosse 31, 115409 Moscow, Russia. *Correspondence e-mail: anthony.teslyuk@grid.kiae.ru, ivan.vartaniants@desy.de

Modern X-ray free-electron lasers (XFELs) operating at high repetition rates produce a tremendous amount of data. It is a great challenge to classify this information and reduce the initial data set to a manageable size for further analysis. Here an approach for classification of diffraction patterns measured in prototypical diffract-and-destroy single-particle imaging experiments at XFELs is presented. It is proposed that the data are classified on the basis of a set of parameters that take into account the underlying diffraction physics and specific relations between the real-space structure of a particle and its reciprocal-space intensity distribution. The approach is demonstrated by applying principal component analysis and support vector machine algorithms to the simulated and measured X-ray data sets.

1. Introduction

Conventional X-ray studies on small biological samples are generally limited for two major reasons. First, well established X-ray crystallographic methods are applicable only to sufficiently large crystals and these cannot be crystallized for all systems of interest (Drenth, 2007). Second, in X-ray imaging of non-crystalline biological samples radiation damage limits achievable resolution to a few tens of nanometres (Henderson, 1995; Howells *et al.*, 2009). The single-particle diffractive imaging approach (Gaffney & Chapman, 2007; Aquila *et al.*, 2015) may allow the second limitation to be overcome and increase the resolution of biological objects to the sub-nanometre range in X-ray experiments at X-ray free-electron lasers (XFELs).

High-power XFELs (Emma *et al.*, 2010; Ishikawa *et al.*, 2012; Altarelli *et al.*, 2007) with a femtosecond pulse duration can be used to perform structure determination experiments on single particles (Gaffney & Chapman, 2007; Seibert *et al.*, 2011; Mancuso *et al.*, 2010). In these experiments, reproducible particles in random orientations are injected into the XFEL beam [see Fig. 1(*a*)]. If X-ray diffraction patterns can be measured before there is any notable radiation damage to the particles (Neutze *et al.*, 2000; Lorentz *et al.*, 2012; Gorobtsov *et al.*, 2015) then the three-dimensional structure of a particle can be reconstructed from these patterns. Such a 'diffract and destroy' approach has been successfully applied to study the structure of protein nanocrystals (Chapman *et al.*, 2011; Boutet *et al.*, 2012) using classical crystallography approaches for Bragg peaks analysis.



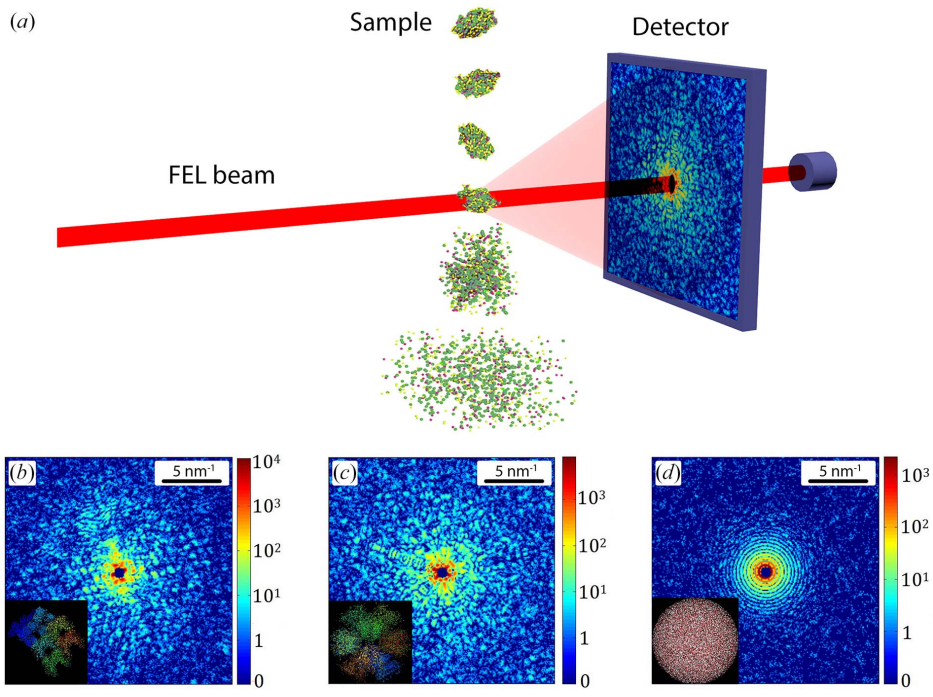


Figure 1
(*a*) Generic scheme of the SPI experiment. Typical simulated diffraction patterns for different particles: (*b*) BTV protein, (*c*) adenovirus protein, (*d*) water droplet.

Single-particle diffractive imaging, in fact, requires the measurement of a large number of diffraction patterns from particles in different orientations, in order to sufficiently sample the three-dimensional intensity distribution in reciprocal space. To determine this three-dimensional intensity distribution a large measured data set should go through a few steps of preprocessing (Gaffney & Chapman, 2007). These preliminary steps include an image classification procedure and orientation analysis. The latter has been discussed in detail elsewhere (Loh & Elser, 2009; Fung *et al.*, 2009; Yefanov & Vartanyants, 2013) and in this paper we focus on classification algorithms.

The necessity of image classification arises owing to the technical aspects of single-particle X-ray diffraction experiments where not all of the measured diffraction patterns contain useful information about the object of interest. Most of the measured images contain only background scattering or blank frames, some may contain diffraction from water droplets or even from some contaminant. In single-particle experiments it is also desirable to distinguish images that originate from multiple particles (Hantke *et al.*, 2014). While these data cannot be used in a classical single-particle imaging (SPI) technique, they could be treated by the X-ray cross-correlation analysis (XCCA) approach (Kurta *et al.*, 2013*a*; Pedrini *et al.*, 2013; Saldin *et al.*, 2011) along with single-particle hits, provided that the inter-particle interference contribution from multiple particles can be neglected. In this way one can substantially increase the amount of useful data measured in imaging experiments at XFELs.

It is, therefore, necessary to classify the measured images before applying an orientation determination procedure, and use only those ones that originate from the particle of interest. While it is possible to perform such classification manually, this would be a very time-consuming procedure, considering that the expected number of measured diffraction patterns in a typical single-particle imaging experiment can be of the order 10^5–10^6 or even larger. Significant reduction of the initial data set can be performed online during the X-ray experiment; for example, using the real-time image rejection based on the signal measured by an ion time-of-flight spectrometer (Andreasson *et al.*, 2014). However, further classification is required to reach the input data quality required for the subsequent orientation determination procedure. All of this motivates the development of efficient computational methods for diffraction pattern classification.

Recently, computational methods for single-particle data classification were proposed that are based on principal component analysis (PCA) or spectral clustering (Yoon *et al.*, 2011), and also on particle size filters determined *via* image autocorrelation functions (Andreasson *et al.*, 2014). Practically, it is very likely that robust data classification will be achieved by combining several different approaches; therefore, the development of novel methods of data classification and extension of already existing ones is an important task.

In image processing, various methods for data classification are based on the extraction and analysis of a so-called 'feature'. Such methods are used in face recognition (Yang *et al.*, 2004), computer vision (Viola & Jones, 2001), linguistics (Sebastiani, 2002) and data mining (Berkhin, 2006). The feature itself represents a small set of parameters which encompass the most valuable information about the object of interest that can be used for data classification. In this paper, we present an approach for X-ray diffraction patterns classification based on PCA (Jolliffe, 2002) and support vector machine (SVM) (Cortes & Vapnik, 1995) algorithms. The basic idea of our approach is that the feature vector can be composed of parameters that take into account the underlying diffraction physics and relations between the real-space structure of a particle and its reciprocal-space intensity distribution.

The paper is organized as follows: in §2 we provide a theoretical basis for our approach of image compression and feature extraction; in §3 we describe the two data sets that were used in the development of our approach; in §4 we demonstrate the results of diffraction pattern classification using PCA and SVM algorithms, and complete the paper with a summary.

2. Theoretical background

In this section, we present a method for feature extraction from the measured diffraction patterns. We propose to parametrize the feature vector in terms of the Fourier components (FCs) of the intensity cross-correlation functions (CCFs). We show that these functions provide useful relations between the electron density of a particle and the scattered intensity distribution; in particular cases leading to very compact features. We also describe a feature vector calculation procedure.

2.1. X-ray cross-correlation analysis of scattered intensities

Kinematically scattered intensity $I_{\psi_i}(\mathbf{q})$ at the momentum transfer $\mathbf{q}$ from a single particle in an arbitrary orientation ψ_i is related to its electron density $\rho_{\psi_i}(\mathbf{r})$ as follows (Als-Nielsen & McMorrow, 2011),

$$I_{\psi_i}(\mathbf{q}) = \left| \int \rho_{\psi_i}(\mathbf{r}) \exp(i\mathbf{qr})\, \mathrm{d}\mathbf{r} \right|^2, \tag{1}$$

where $\mathbf{r}$ is a real-space coordinate. The intensity distribution $I_{\psi_i}(\mathbf{q})$ measured in the experiment on a two-dimensional detector represents a section of reciprocal space by the Ewald sphere. It can be expressed in the polar coordinate system of the detector $\mathbf{q} = (q, \varphi)$ $(0 < \varphi \leq 2\pi)$, in the form of angular Fourier series:

$$I_{\psi_i}(q, \varphi) = \sum_{n=-\infty}^{\infty} I^n_{q,\psi_i} \exp(in\varphi), \tag{2}$$

$$I^n_{q,\psi_i} = \frac{1}{2\pi} \int_0^{2\pi} I_{\psi_i}(q, \varphi) \exp(-in\varphi)\, \mathrm{d}\varphi, \tag{3}$$

where I^n_{q,ψ_i} are the FCs of $I_{\psi_i}(q, \varphi)$. The set of the FCs I^n_{q,ψ_i} calculated for all possible resolution rings q completely determines the scattered intensity $I_{\psi_i}(q, \varphi)$ measured on a given diffraction pattern. At the same time these FCs are defined by the unique electron density distribution $\rho_{\psi_i}(\mathbf{r})$ of the particle, which makes them attractive parameters for identification of particles with different structures.

Angular X-ray cross-correlation analysis (Wochner *et al.*, 2009) is a technique that allows statistical studies of the scattered intensities and provides convenient means of extraction of the FCs I^n_q of intensity (Altarelli *et al.*, 2010; Kurta *et al.*, 2013*b*). The basic component of XCCA is an angular intensity CCF. The two-point CCFs can be defined on two resolution rings q_1 and q_2 (Kam, 1977; Altarelli *et al.*, 2010),

$$C_{\psi_i}(q_1, q_2, \Delta) = \left\langle I_{\psi_i}(q_1, \varphi) I_{\psi_i}(q_2, \varphi + \Delta) \right\rangle_\varphi, \tag{4}$$

where $0 \leq \Delta \leq 2\pi$ is the angular coordinate and $\langle \ldots \rangle_\varphi$ denotes the average over the angle φ. It can be directly shown (Altarelli *et al.*, 2010) that the FCs $C^n_{q_1,q_2,\psi_i}$ of the CCF $C_{\psi_i}(q_1, q_2, \Delta)$ for $n \neq 0$ are defined by the FCs of the scattered intensity,

$$C_{\psi_i}(q_1, q_2, \Delta) = \sum_{n=-\infty}^{\infty} C^n_{q_1,q_2,\psi_i} \exp(in\Delta), \tag{5}$$

$$C^n_{q_1,q_2,\psi_i} = \frac{1}{2\pi} \int_0^{2\pi} C_{\psi_i}(q_1, q_2, \Delta) \exp(-in\Delta)\mathrm{d}\Delta, \tag{6}$$

$$C^n_{q_1,q_2,\psi_i} = I^{n*}_{q_1,\psi_i} I^n_{q_2,\psi_i}. \tag{7}$$

In the particular case of cross-correlation on the same resolution ring $q_1 = q_2 = q$, equations (4) and (7) reduce to

$$C_{\psi_i}(q, \Delta) = \left\langle I_{\psi_i}(q, \varphi) I_{\psi_i}(q, \varphi + \Delta) \right\rangle_\varphi \tag{8}$$

and

$$C^n_{q,\psi_i} = \left| I^n_{q,\psi_i} \right|^2, \tag{9}$$

respectively. In this case the magnitudes of the FCs of intensity can be directly determined as $|I^n_{q,\psi_i}| = (C^n_{q,\psi_i})^{1/2}$. The advantage of applying the CCF here is that the undesirable experimental factors (*e.g.* the presence of gaps between detector tiles or masked/missing data on the measured diffraction patterns) can be eliminated from the data analysis.

Analysis of the FCs C^n_{q,ψ_i} has shown that for two-dimensional particles with a certain symmetry the FCs of specific orders n should be predominant (Altarelli *et al.*, 2010; Kurta *et al.*, 2012). For example, scattering from a two-dimensional particle with a fivefold rotational symmetry axis parallel to the incoming beam would lead, in the case of a curved Ewald sphere, to dominance of FCs of the orders n that are multiples of five ($n \bmod 5 = 0$), *i.e.* $n = 5, 10, 15, \ldots$. For a flat Ewald sphere, the Friedel symmetry allows only even orders of the FCs and one would observe only FCs of the orders $n = 10, 20, \ldots$. In the case of a particle without any pronounced symmetry, there will be no dominating components. Therefore, simple analysis of the Fourier spectra of the CCF could help to distinguish particles with certain rotational symmetry and without symmetry. The feature vector in this case can be represented by a small number of FCs of the CCF.

In the case of three-dimensional particles the situation is more complicated because the X-ray beam scatters from a particle in different orientations ψ_i in each XFEL snap-shot that leads in each case to a different spectrum C^n_{q,ψ_i}. If particle orientations are uniformly distributed, the spectra averaged over many diffraction patterns converge to an average $\langle C^n_{q,\psi_i} \rangle_{\psi_i}$ that can be used to distinguish between different types of particles. Interestingly, even in the case of simultaneous scattering from a few identical particles in different orientations ψ_i, the average $\langle C^n_{q,\psi_i} \rangle_{\psi_i}$ converges to the spectrum of just a single particle, provided that the inter-particle interference contribution can be neglected (Kurta *et al.*, 2012, 2013*a*,*b*). Importantly, the q-dependence of the FCs $\langle C^n_{q,\psi_i} \rangle_{\psi_i}$ has a characteristic profile defined by a particular structure of a particle (Kurta *et al.*, 2012, 2013*b*). This brings us to conclude that, even in the case of three-dimensional particles, the FCs C^n_{q,ψ_i} contain specific information that encodes the underlying particle structure. We utilize these properties of the CCFs and their Fourier spectra in our approach to distinguish diffraction patterns originating from different sources of scattering.

2.2. Feature vector for diffraction pattern classification

In our image classification approach we parametrize the feature vector F in terms of the FCs of the CCF introduced in the previous section. Here we describe a general procedure for feature vector construction. Some practical details of data preprocessing are discussed in Appendix A.

Owing to the specific features of an SPI experiment, information in a diffraction pattern is distributed non-uniformly. A beamstop may cover a central part of the diffraction pattern and gaps between detector tiles lead to an additional loss of information in different parts of the measured patterns. Also, the scattered intensity decreases rapidly as a function of q, and at higher momentum transfer values the pattern contains rare photon counts. Therefore, in image classification it is reasonable to analyze the intensity distribution only within a certain region of interest (ROI) in the form of an annulus $q_{\min} \leq q \leq q_{\max}$, where the measured signal is mostly informative (see Fig. 2 and Appendix A).

In the selected ROI the diffraction patterns typically have noticeable differences. To parametrize these differences quantitatively we applied to each image[1] the CCF $C(q, \Delta)$ defined in equation (8). Determined in such a way, the two-dimensional CCF $C(q, \Delta)$ was subsequently averaged over the q-range within the defined ROI in order to reduce the number of parameters:

$$C(\Delta) = \langle C(q, \Delta) \rangle_q. \tag{10}$$

The result calculated for each diffraction pattern was normalized by the angular averaged squared intensity also averaged over ROI:

$$\bar{C}(\Delta) = \frac{C(\Delta)}{\langle I(q, \varphi)^2 \rangle_{\varphi,q}} = \frac{C(\Delta)}{C(0)}. \tag{11}$$

Such normalization is necessary to reduce the effect of intensity fluctuations from one diffraction pattern to another, owing to intensity fluctuations of the incoming XFEL beam as well as particle position in the focused XFEL beam.

In a similar way to the CCF $C(q, \Delta)$, the function $\bar{C}(\Delta)$ defined in equation (11) can be expanded into a Fourier cosine series,

$$\bar{C}(\Delta) = 2 \sum_{n=1}^{\infty} \bar{C}^n \cos(n\Delta), \tag{12}$$

$$\bar{C}^n = \frac{1}{\pi} \int_0^{\pi} \bar{C}(\Delta) \cos(n\Delta)\, \mathrm{d}\Delta, \tag{13}$$

where $\bar{C}^n$ are the FCs of $\bar{C}(\Delta)$. The FCs $\bar{C}^n$ constitute a compact set of parameters that carry information on angular features of the diffraction patterns. Typically, a limited number m (that is much smaller than the total number of pixels on a detector) of nonzero FCs contributes to the spectrum $\bar{C}^n$, depending on the particle structure and experimental conditions (Kurta *et al.*, 2013*a*,*c*). This helps to reduce dimensionality of the feature vector **F** and to speed up the data analysis.

[1] In this sub-section, similar to the previous one, all quantities are defined for a particular particle orientation ψ_i, with the subscript ψ_i omitted for brevity.

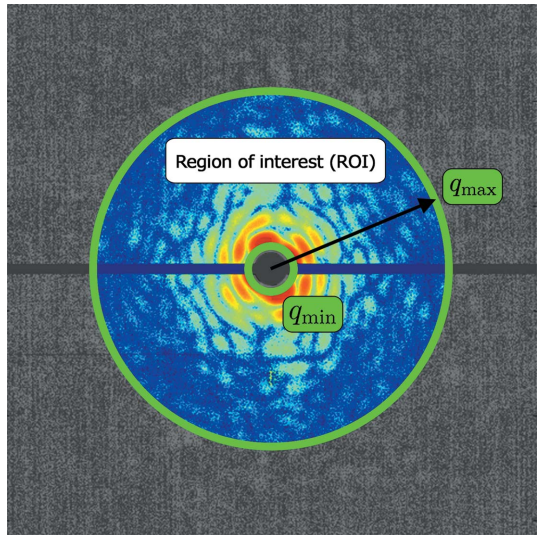


Figure 2
The region of interest (ROI) of a diffraction pattern, defined in the form of an annulus $q_{\min} < q < q_{\max}$.

Additional information about a particle can be obtained from the normalized q-dependence of $C(q, 0) = \langle I(q, \varphi)^2 \rangle_{\varphi}$ [see equation (8)] within the defined ROI:

$$\bar{C}_q = \frac{C(q, 0)}{\langle I(q, \varphi) \rangle_{\varphi}^2} = \frac{\langle I(q, \varphi)^2 \rangle_{\varphi}}{\langle I(q, \varphi) \rangle_{\varphi}^2}. \tag{14}$$

Two subsets of parameters defined by $\bar{C}^n$ and $\bar{C}_q$ can differ significantly in terms of the corresponding values of variance determined for each subset. Therefore, the effect of one of the subsets may be diminished compared with another, while performing classification by means of PCA. To avoid such an undesirable situation one can linearly transform two subsets α and β using the weights a and b expressed by the respective values of standard deviation,

$$a = 1/\sigma_{\alpha}, \qquad b = 1/\sigma_{\beta}. \tag{15}$$

Here the values of standard deviations σ_{α} and σ_{β} are defined as

$$\sigma_{\alpha} = \left[\frac{1}{N_m} \sum_{n=1}^{N_m} \left(\sigma_{\alpha}^n\right)^2 \right]^{1/2}, \quad \sigma_{\beta} = \left[\frac{1}{N_q} \sum_{q=q_{\min}}^{q_{\max}} \left(\sigma_{\beta}^q\right)^2 \right]^{1/2}, \tag{16}$$

where

$$\sigma_{\alpha}^n = \frac{1}{M} \sum_{i=1}^{M} \left(\bar{C}_{\psi_i}^n - \langle \bar{C}_{\psi_i}^n \rangle_{\psi_i} \right)^2$$

and

$$\sigma_{\beta}^q = \frac{1}{M} \sum_{i=1}^{M} \left(\bar{C}_{q,\psi_i} - \langle \bar{C}_{q,\psi_i} \rangle_{\psi_i} \right)^2.$$

In (16), summation is performed over N_m FCs and N_q values in the range of q from $q_{\min}$ to $q_{\max}$. The sum over different orientations ψ_i is practically realised as a sum over M measured diffraction patterns. The weighted parameters $\bar{C}^n$ and $\bar{C}_q$ can then be used to construct the feature vector **F** of a diffraction pattern as

$$\mathbf{F} = \left(a\bar{C}^1, \ldots, a\bar{C}^m, b\bar{C}_{q_{\min}}, \ldots, b\bar{C}_{q_{\max}}\right). \quad (17)$$

The resulting feature vector is determined by the particle structure and its orientation with respect to the incoming beam direction.

3. Data description

Here we describe the two data sets that were used in the development and testing of our image classification algorithms.

3.1. Simulated data

The simulated data set consists of diffraction patterns calculated for three types of particles: adenovirus 2/12 penton base chimera [entry 2c6s in the Protein Data Bank (PDB)] (Zubieta *et al.*, 2006) referred below as adenovirus protein; an assembly of VP3 and VP7 proteins from the bluetongue virus core (entry 2btv in the PDB) (Grimes *et al.*, 1998) referred below as BTV protein; and a 10 nm-diameter droplet of water. The adenovirus protein has a hexagonal symmetry, while the BTV protein is asymmetric. The particles were chosen to have a comparable size to complicate their classification. For each particle type 1000 images were generated. The simulated data set was divided into two groups: 100 training images for which the particle type was known and the remaining 2900 images for classification with our algorithm.

In X-ray scattering simulations a detector was assumed to be located at a distance of 100 mm from the sample and to have a size of 100 mm × 100 mm and 224 × 224 pixels in total. The X-ray wavelength was considered to be 0.3 nm. The angular speckle size was 0.06 rad for this configuration. The beam incident on the sample was assumed to be Gaussian with full width at half-maximum of 150 nm and a fluence of 10^7 photons nm^{-2}, which is a typical fluence for a focused incident beam at the XFEL facilities (Emma *et al.*, 2010; Ishikawa *et al.*, 2012). Poisson noise and a beamstop of 10 pixels in diameter (about 0.74 of a speckle) were applied to generated patterns. Simulations were performed using *MOLTRANS* code.

Typical diffraction patterns for different samples are shown in Figs. 1(*b*)–1(*d*). While water droplets produce images that are quite distinctive and can be easily identified, patterns from the other two particles look similar. Therefore, the main objective was to distinguish diffraction patterns originating from the bluetongue virus core and the adenovirus penton base chimera.

3.2. Experimental data

In addition to the simulated data set, we used the experimental data from Kassemeyer *et al.* (2012) which were deposited in the Coherent X-ray Imaging Data Bank (Maia, 2012). The measurements were performed at the Atomic, Molecular and Optical Science beamline (Bozek, 2009) at the LCLS using the CFEL-ASG Multi-Purpose (CAMP) endstation (Strüder *et al.*, 2010). In this experiment, diffraction patterns from *Paramecium bursarium chlorella* virus (PBCV-1) (Van Etten *et al.*, 1983) and bacteriophage T4 (Kassemeyer *et al.*, 2012) were measured. As compared with the simulated patterns, the experimental data set is more challenging for classification. Besides the effect of experimental factors, for example, fluctuations of the incident beam position and fluence, the samples themselves can be different: particles could be coated by the solvent or multiple particles could be present in the beam. We considered for classification a random mixture of 532 patterns from the PBCV-1 and 964 patterns from the bacteriophage T4. The experimental data set was also divided into two groups: one with 100 training images and the other with 1396 images for classification with our algorithm.

4. Results

Here we present the results of image classification for the simulated and the experimental data sets. For each image, the feature vector was determined according to equation (17) with the FCs $\bar{C}^n$ [equation (13)] up to the maximum order $m = 50$ and parameters $\bar{C}_q$ [equation (14)] were determined with a one pixel sampling rate within the defined ROI, $q_{\min} \leq q \leq q_{\max}$. The feature vectors were used as input data in PCA and SVM algorithms to perform data clustering.

4.1. Classification of X-ray images by means of PCA

Briefly, PCA is an orthogonal linear transformation that converts the data to a new coordinate system defined in terms of the so-called principal components (Jolliffe, 2002). The principal components are the eigenvectors of the data covariance matrix, arranged in descending order of the corresponding eigenvalues. In the present context, the data covariance matrix is constructed using the feature vectors of the diffraction images (see Appendix *B* for details). One of the main PCA features is that the principal component space constructed using the first n components has the largest possible variance among any possible n-dimensional orthonormal bases. Therefore, the feature vectors projected onto a two-dimensional coordinate space defined by the first two principal components should account for as much of the variability in the data as possible. In this way we transform the input set of feature vectors to a new set of parameters that can be visualized as points on a two-dimensional plane corresponding to different images, which can be subsequently classified.

The feature vectors determined for the simulated data set were projected onto a two-dimensional plane defined by the first two principal components (PC1, PC2). The results of such a transformation are presented in Fig. 3. First, we applied PCA to the training data set to see if the images corresponding to different types of particles can be visually separated (orange and red dots). The images corresponding to water droplets fall into one distant group that can be easily identified (black dots). Then, the rest of the data set was classified by PCA, showing clustering of the data into two different sets of points outlined by the training set (blue and green dots). To establish

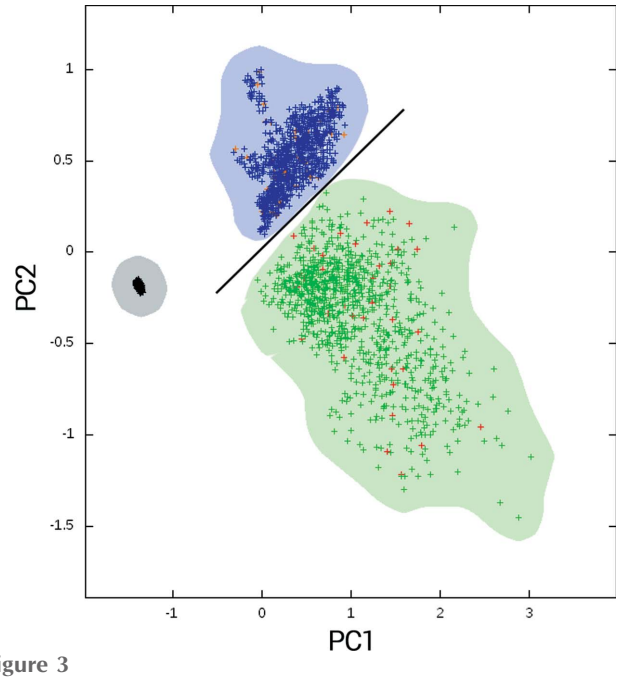


Figure 3
Results of PCA for the simulated data set. Every dot corresponds to image coordinates in the plane formed by the first two principal components (PC1, PC2). Orange dots correspond to adenovirus protein images and red dots correspond to BTV protein diffraction patterns that were used for algorithm training. Blue and green dots correspond to images of the adenovirus protein and the BTV protein, respectively, which were classified by the PCA algorithm. Black dots correspond to diffraction patterns from water droplets.

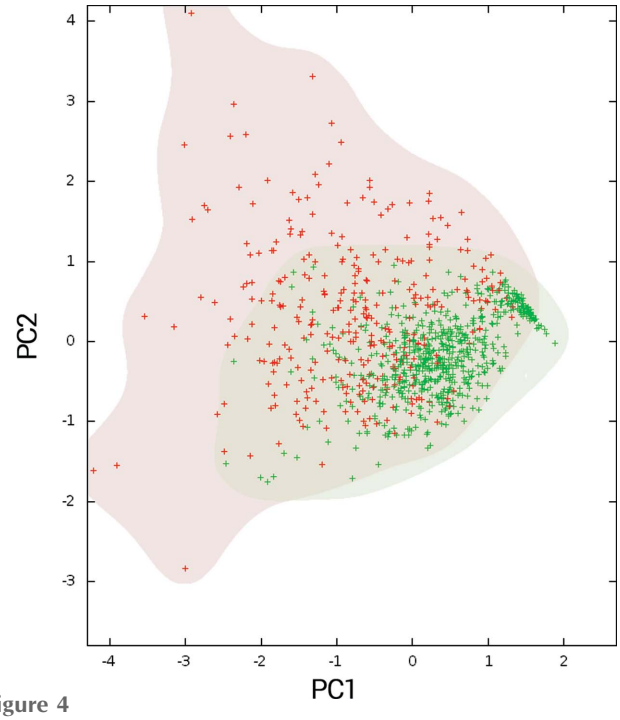


Figure 4
Results of PCA for the experimental data set. The dots correspond to two types of images, measured from PBCV-1 (red dots) and bacteriophage T4 (green dots) particles. Patterns are colored according to the final SVM result (see text). Clustering of the data is not sufficient for reliable image classification.

a classification rule one needs to draw a dividing line between different clusters on the PC1–PC2 plane. If different clusters do not overlap with each other on the plane (as it is in our case, see Fig. 3), one can readily set up a separating line between them. This line may be not optimal as it is understood in classification methods like SVM but, in our case, it gives 100% accuracy of classification for the entire data set. Notice that small modifications of the feature vector [equation (17)] may lead to a significant overlap of two clusters on the PC1–PC2 plane. In that case, clusters could be visually separated in three-dimensional space (PC1–PC2–PC3) or on the PC1–PC3 plane. For the particular data set considered here, clusters do not overlap in PC1–PC2–PC3 space. This is a problem of the PCA method because components of the largest possible variance do not strictly correspond to components of best separation. We believe that successful classification of the simulated data set was possible due to the appropriate selection of the feature extraction method.

4.2. Classification of X-ray images by means of SVM

A similar clustering procedure based on PCA was applied to the experimental data set, with the results presented in Fig. 4. As the experimental data set is more complex compared with the simulated one, the same procedure fails to achieve the desired level of clustering. As one can see, visual separation of images corresponding to different types of particles has not been achieved in this case (see Fig. 4).

Next, we applied the SVM algorithm in an attempt to classify the experimental data. Generally, SVM constructs a hypersurface in the feature vector space which can be used to separate different image types (Cortes & Vapnik, 1995). In the present case of two different types of particles we applied linear SVM, where the optimal hyperplane can be determined by maximizing the distance between the points associated with different types of particles. As SVM employs learning algorithms to construct such a hyperplane, it requires a training data set with known image types, which was provided as in the case of clustering with PCA. In the case of a few particles, the multiclass SVM can be implemented by recursive application of SVM to a certain class of images against all other image types. First, the SVM-based classification was verified on the simulated data set (Fig. 5). Our results show that SVM provides better separation of different clusters than PCA. The results of classification of the experimental data using SVM are presented in Fig. 6, where the determined hyperplane crosses the origin of the horizontal axis and is aligned parallel to the vertical axis (perpendicular to the image plane). The distance from a point to the separating hyperplane (horizontal coordinate) could be considered as a probability of classification (class score): the longer the distance, the higher probability of a point to belong to a certain class. Probability is determined by Platt scaling (Platt, 1999):

$$P(y = 1/x) = 1/(1 + \exp x), \tag{18}$$

where x is a class score. According to this definition, points that have a class score close to zero could not be classified

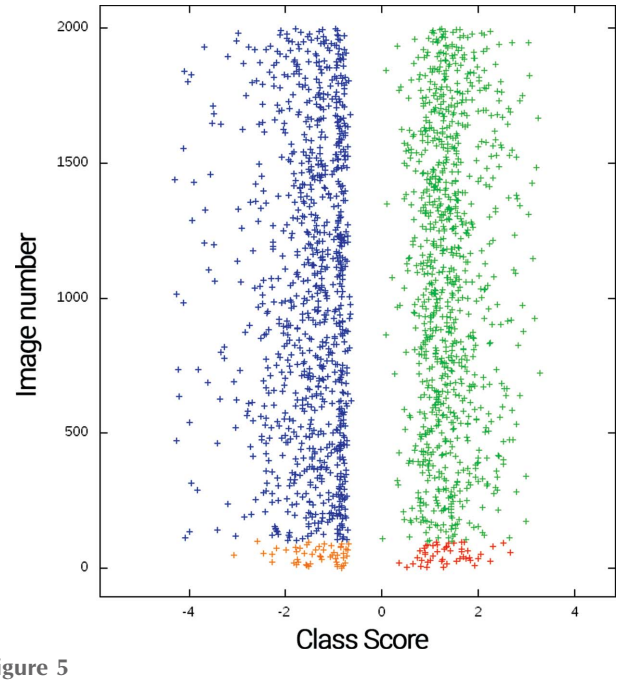


Figure 5
Results of the SVM classification for the simulated data set for adenovirus protein and BTV protein diffraction patterns. The horizontal coordinate corresponds to the probability of a diffraction pattern being related to a certain type of a particle. The vertical coordinate corresponds to the pattern number. Blue dots correspond to adenovirus protein images and green dots correspond to BTV protein diffraction patterns. Yellow (adenovirus) and red (BTV) dots were used for training.

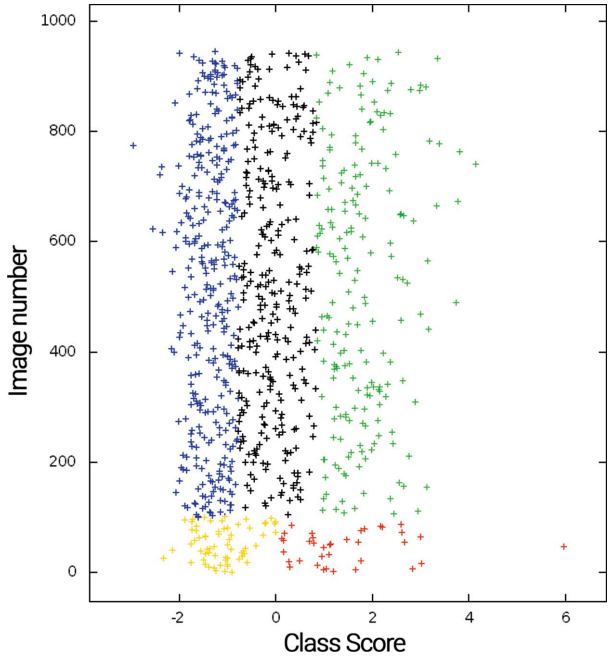


Figure 6
Results of the SVM classification for the experimental data set. The horizontal coordinate corresponds to the probability of a diffraction pattern being related to a certain type of a particle. The vertical coordinate corresponds to the pattern number. Blue dots correspond to bacteriophage T4 and green dots correspond to PBCV-1 particles. Black dots correspond to patterns with probability of correct classification below 75%. Yellow (bacteriophage T4) and red (PBCV-1) dots were used for training.

reliably, *i.e.* have smaller probability to belong to a certain cluster. For the experimental data set, 87% of images with the probability of correct classification above 75% were properly classified.

5. Summary

Future XFELs will operate at MHz repetition rates (Altarelli *et al.*, 2007) and produce a tremendous amount of data in short periods of time. Here we have presented an approach that should facilitate classification of diffraction patterns measured in prototypical diffract-and-destroy single-particle imaging experiments at XFELs. It is aimed to selectively extract the necessary information and to reduce the initial data set to a manageable size for further analysis. The proposed approach is based on the image feature extraction, parametrized in terms of the Fourier spectra of the angular intensity cross-correlation functions that maintain relations between the real space structure of a particle and its reciprocal space intensity distribution. We demonstrated our approach by applying PCA and SVM algorithms to simulated and measured X-ray data sets. While both methods demonstrate accurate clustering of the simulated data, only the SVM algorithm allowed us to classify different biological species in the experimental data set. We believe that the success achieved in data clustering is largely determined by the specific parametrization of the image feature. Such feature parametrization is the central result of the paper and can be adopted for other algorithms of X-ray diffraction data classification.

APPENDIX *A*
Data preprocessing

A1. Simulated data

In the case of the simulated data set, the incident intensity and the particle position in the beam was the same for each simulated image, as well as the position of the pattern center on the detector. The ROI ($q_{\rm min}$, $q_{\rm max}$) (see Fig. 2) was chosen in such a way that $q_{\rm min}$ was a few pixels larger than the size of the central beamstop and $q_{\rm max}$ corresponds to the momentum transfer where the average scattered intensity was twice as high as the signal near the image boundaries, where we believe it contains mostly noise and rare random photon counts. The determined ROI was kept the same for all images to provide the same length of the feature vector **F** [equation (17)].

A2. Experimental data

Compared with the simulated data set, the experimental one contains a few points that have to be properly treated:

(*a*) The center of a diffraction pattern was determined for each image prior to classification. It was also necessary to adjust the value of $q_{\rm min}$ to completely cover the beamstop area for all determined positions of the pattern center. A natural choice for $q_{\rm min}$ was its maximum value determined for the whole data set.

(*b*) Some of the measured patterns were supersaturated and/or contain parasitic scattering. In such cases bad pixels were discarded from the analysis by masking.

(*c*) A number of diffraction patterns contained only background scattering or blank images. To filter out such images, we prepared a training data set composed of such images and applied the classification algorithm to divide the entire data set into two classes of images: informative and non-informative. Further classification was performed only for the informative images.

APPENDIX *B*
Principal component analysis calculations

To perform PCA, we first constructed a data matrix, $A = ||a_{ij}||$, in which indices i and j specify the ith image in the data set and the jth component of the feature vector **F** [equation (17)] of the corresponding image. Then, the eigenvectors of the data covariance matrix, *i.e.* the principal components, can be evaluated. The fastest way to do this is first to column-center the matrix A to obtain the matrix $\bar{A} = ||\bar{a}_{ij}||$:

$$\bar{a}_{ij} = a_{ij} - \frac{1}{N}\sum_{k=0}^{N} a_{ki}. \tag{19}$$

The principal components sought are the eigenvectors of the covariance matrix $\bar{A}^T\bar{A}$. Using a singular value decomposition one can write

$$\bar{A} = U\Sigma V^T, \tag{20}$$

where U and V are unitary matrices and Σ is a diagonal. Finally, using equation (20) we obtain

$$\bar{A}^T\bar{A} = V(\Sigma^T\Sigma)V^T. \tag{21}$$

Because $\Sigma^T\Sigma$ is a diagonal matrix, the columns of V are the principal component vectors we are looking for.

Acknowledgements

Numerical simulations were performed at supercomputing resources in the NRC 'Kurchatov Institute', which are supported as the centre for collective usage (project RFMEFI62114X0006, funded by Ministry of Science and Education of Russia). This work was partially supported by the Virtual Institute VH-VI-403 of the Helmholtz Association. Discussions with and support of this project by E. Weckert are greatly acknowledged.

References

Als-Nielsen, J. & McMorrow, D. (2011). *Elements of Modern X-ray Physics*, 2nd ed. New York: Wiley.
Altarelli, M. *et al.* (2007). *The European X-ray Free-Electron Laser.* Technical Design Report 2006-097. DESY, Hamburg, Germany.
Altarelli, M., Kurta, R. P. & Vartanyants, I. A. (2010). *Phys. Rev. B*, **82**, 104207. [Erratum: (2012). *Phys. Rev. B*, **86**, 179904.]
Andreasson, J. *et al.* (2014). *Opt. Express*, **22**, 2497–2510.
Aquila, A. *et al.* (2015). *Struct. Dyn.* **2**, 041701.
Berkhin, P. (2006). *Grouping Multidimensional Data, Recent Advances in Clustering*, pp. 25–71. Berlin/Heidelberg: Springer.
Boutet, S. *et al.* (2012). *Science*, **337**, 362–364.
Bozek, J. D. (2009). *Eur. Phys. J. Spec. Top.* **169**, 129–132.
Chapman, H. N. *et al.* (2011). *Nature (London)*, **470**, 73–77.
Cortes, C. & Vapnik, V. (1995). *Mach. Learn.* **20**, 273–297.
Drenth, J. (2007). *Principles of Protein X-ray Crystallography*. Berlin: Springer.
Emma, P. *et al.* (2010). *Nat. Photon.* **4**, 641–647.
Fung, R., Shneerson, V., Saldin, D. K. & Ourmazd, A. (2009). *Nat. Phys.* **5**, 64–67.
Gaffney, K. J. & Chapman, H. N. (2007). *Science*, **316**, 1444–1448.
Gorobtsov, O. Y., Lorenz, U., Kabachnik, N. M. & Vartanyants, I. A. (2015). *Phys. Rev. E*, **91**, 062712.
Grimes, J. M., Burroughs, J. N., Gouet, P., Diprose, J. M., Malby, R., Zientara, S., Mertens, P. P. & Stuart, D. I. (1998). *Nature (London)*, **395**, 470–478.
Hantke, M. F. *et al.* (2014). *Nat. Photon.* **8**, 943–949.
Henderson, R. (1995). *Q. Rev. Biophys.* **28**, 171–193.
Howells, M. R., Beetz, T., Chapman, H. N., Cui, C., Holton, J. M., Jacobsen, C. J., Kirz, J., Lima, E., Marchesini, S., Miao, H., Sayre, D., Shapiro, D. A., Spence, J. C. H. & Starodub, D. (2009). *J. Electron Spectrosc. Relat. Phenom.* **170**, 4–12.
Ishikawa, T. *et al.* (2012). *Nat. Photon.* **6**, 540–544.
Jolliffe, I. T. (2002). *Principal Component Analysis, Springer Series in Statistics.* New York: Springer-Verlag.
Kam, Z. (1977). *Macromolecules*, **10**, 927–934.
Kassemeyer, S. *et al.* (2012). *Opt. Express*, **20**, 4149–4158.
Kurta, R. P., Altarelli, M. & Vartanyants, I. A. (2013*b*). *Adv. Condens. Matter Phys.* **2013**, 959835.
Kurta, R. P., Altarelli, M., Weckert, E. & Vartanyants, I. A. (2012). *Phys. Rev. B*, **85**, 184204.
Kurta, R. P., Chesnokov, Y., Weckert, E. & Vartanyants, I. A. (2013*c*). *J. Phys. Conf. Ser.* **463**, 012046.
Kurta, R. P., Dronyak, R., Altarelli, M., Weckert, E. & Vartanyants, I. A. (2013*a*). *New J. Phys.* **15**, 013059.
Loh, N. D. & Elser, V. (2009). *Phys. Rev. E*, **80**, 026705.
Lorentz, U., Kabachnik, N. M., Weckert, E. & Vartanyants, I. A. (2012). *Phys. Rev. E*, **86**, 051911.
Maia, F. R. N. C. (2012). *Nat. Methods*, **9**, 854–855.
Mancuso, A. P., Yefanov, O. M. & Vartanyants, I. A. (2010). *J. Biotechnol.* **149**, 229–237.
Neutze, R., Wouts, R., van der Spoel, D., Weckert, E. & Hajdu, J. (2000). *Nature (London)*, **406**, 752–757.
Pedrini, B., Menzel, A., Guizar-Sicairos, M., Guzenko, V. A., Gorelick, S., David, C., Patterson, B. D. & Abela, R. (2013). *Nat. Commun.* **4**, 1647.
Platt, J. C. (1999). *Advances in Large Margin Classifiers*, pp. 61–74. MIT Press.
Saldin, D. K., Poon, H.-C., Schwander, P., Uddin, M. & Schmidt, M. (2011). *Opt. Express*, **19**, 17318.
Sebastiani, F. (2002). *ACM Comput. Surv.* **34**, 1–47.
Seibert, M. M. *et al.* (2011). *Nature (London)*, **470**, 78–81.
Strüder, L. *et al.* (2010). *Nucl. Instrum. Methods Phys. Res. A*, **614**, 483–496.
Van Etten, J. L., Burbank, D. E., Xia, Y. & Meints, R. H. (1983). *Virology*, **126**, 117–125.
Viola, P. & Jones, M. (2001). *Comput. Vis. Pattern Recognit.* **1**, I-511–I-518.
Wochner, P., Gutt, C., Autenrieth, T., Demmer, T., Bugaev, V., Ortiz, A. D., Duri, A., Zontone, F., Grübel, G. & Dosch, H. (2009). *Proc. Natl Acad. Sci. USA*, **106**, 11511–11514.
Yang, J., Zhang, D., Frangi, A. F. & Yang, J. Y. (2004). *IEEE Trans. Pattern Anal. Mach. Intell.* **26**, 131–137.
Yefanov, O. M. & Vartanyants, I. A. (2013). *J. Phys. B*, **46**, 164013.
Yoon, C. H. *et al.* (2011). *Opt. Express*, **19**, 16542.
Zubieta, C., Blanchoin, L. & Cusack, S. (2006). *FEBS J.* **273**, 4336–4345.