# Registering Particle Data Sets Using a Rotation and Translation Invariant Nearest-Neighbor Algorithm 

Nicholas W M Ritchie, Scott Wight, Diana Ortiz-Montalvo, Abigail P Lindstrom


## Original Article

# Registering Particle Data Sets Using a Rotation and Translation Invariant Nearest-Neighbor Algorithm 

Nicholas W. M. Ritchie* ${ }^{(D)}$, Scott Wight ${ }^{(D)}$, Diana Ortiz-Montalvo ${ }^{(D)}$, and Abigail P. Lindstrom ${ }^{(D)}$<br>National Institute of Standards and Technology, Materials Measurement Science Division, Gaithersburg, MD 20899, USA<br>*Corresponding author: Nicholas W. M. Ritchie, E-mail: nicholas.ritchie@nist.gov


#### Abstract

It can be useful to register (or align) two sets of particle data measured from the same physical sample. However, if the two data sets were collected at different translational or rotational offsets, finding the optimal registration can be a challenge. We will present an algorithm that efficiently determines the rotation and translational offset that best registers (in a least-squares sense) the corresponding particles in two or more data sets measured from the same sample. This algorithm can be used to merge two data sets that have been collected on overlapping but otherwise distinct regions on the sample. Alternatively, it can be used to overlay data sets that have been collected on the same sample area to compare replicate data for quality control and measurement efficiency purposes.


Key words: alignment, coordinate registration, optical microscopy, particle analysis, scanning electron microscopy

## Introduction

Optical and electron microscopy of various types are often used to study collections of particles. Examples include forensic applications like gunshot residue and soil comparison and industrial applications like additive manufacturing feed-stock material studies or quality control on powder materials. Usually, the particles are dispersed on a substrate and the substrate is imaged. The particles are identified and the position, size, morphology, and other properties of the particles are measured. In this kind of an analysis, there is a ground truth represented by the particles on the substrate and then there is a measurement of the sample. Ideally, the measurement will accurately reflect the ground truth but there will be disparities in position, size, and all the other measured properties. Measuring the same sample multiple times can provide insight into the characteristics of these disparities. However, to get a really representative understanding of the disparities, it is often necessary to treat each analysis of the sample as a distinct measurement. Often this means repeating each measurement as though it were a totally independent event by taking the sample out of the instrument between measurements. As a result, the orientation and translational offset will vary from analysis to analysis. Often the measured data sets will differ not only in the measured quantities but also in the particles identified. This is particularly true if the particle size approaches the detection limit or the stage is moved such that there are un-analyzed strips bordering the analyzed stage fields.

It is useful therefore to be able to register two (or more) measured particle data sets from the same sample by determining the optimal rotation and translational offsets that best align the corresponding particles on each data set.

Computer science has identified a similar class of problems which it groups under the heading of point cloud registration. Two common applications of point cloud registration are (1) registering separately acquired portions of a 3D scanned object and (2) registering time sequential laser imaging,
detection, and ranging (LIDAR) views of an environment. The second problem has seen particular interest as it serves as a key component of many self-driving car systems. Solutions must be both fast and robust to facilitate real-time feedback. Similar algorithms are also used to register images in a panorama. Visually similar points are identified on multiple images and these points are registered. A recent discussion of point cloud registration algorithms and their shortcoming can be found in Maiseli et al. (2017).

However, the class of point cloud registration problems is similar but subtly different from our problem. The most significant difference is that, in our problem, the two data sets each represent separate views of the same underlying particle distribution. In our problem, there is a baseline reality represented by the real positions of the particles, whereas with LIDAR, for example, the measurement points are unlikely to be placed on the same points on the measured surface between data sets. While not all particles are represented in each data set, we assume that a fraction ( $>10 \%$ ) is represented in both. This allows us to potentially assign true one-to-one correspondences between some of the points in both data sets. In contrast, two LIDAR scans of a region are not likely to represent the same points, so there is no natural one-to-one correspondence between points in the data sets. This one-to-one correspondence is what makes our problem interesting and unique. If we can figure out which particles correspond, then we have additional data that is not available in generic point cloud registration problems.

The classic algorithm for point cloud registration is the iterative closest point (ICP) algorithm (Besl \& McKay, 1992). ICP is frequently used to refine the alignment of multidimensional data sets derived from LIDAR or other 3D scanners. It uses a nearest-neighbor algorithm to identify correspondences between close points in the two data sets and singular value decomposition to compute a matrix that transforms the points in one data set towards the points in

[^0]the other. It can be readily implemented in massively parallel systems like graphics processing units, making it suitable for large data sets with real-time demands like with automated driving systems. The ICP algorithm has been used to align timesequential LIDAR views of the surrounding environment. However, the ICP algorithm has a limited range of angular and translational convergence and so must be seeded with a good initial estimate of these parameters. Otherwise, it can fall into a false, local minimum that does not reflect the desired global minimum. A refinement of the ICP algorithm that uses a Levenberg-Marquart nonlinear optimization (Moré, 1978) has been proposed (Fitzgibbon, 2003). The ICP algorithm has a weakness in that it will not necessarily converge to the correct solution if the initial guess is too far from the optimal result. Thus, the ICP algorithm is useful for refining preexisting solutions but not as useful for raw data sets with arbitrary transformations.

Our goal, therefore, is to develop an algorithm that will allow us to assign likely one-to-one correspondences between particles in the two data sets. The algorithm is implemented in the Julia ${ }^{1}$ language (Bezanson et al., 2017). To make this algorithm as generally applicable as possible, we would like to use only the particle location data (stage coordinates) and not other measured particle properties like size, morphology, or composition. We take as our inspiration, the way that our eye can align point clouds by discerning distinct groupings of particles that it can identify in both data sets. These distinct groupings are then used to sequentially pin degrees-of-freedom until we have developed a global solution. The key is the mathematical notion of invariants.

## Algorithm

## Invariants

An invariant is a property of a mathematical object that remains unchanged when a certain operation is performed on the object. Common types of invariants include rotational invariants, translational invariants, mirror reflection invariants, and scale invariants. The key to performing the alignment is to leverage properties of the measured data sets that we know are invariant under the types of operations that have been performed on the data, namely rotation and translation. These properties can be compared between data sets to identify features common to both data sets. These common features can then be exploited to determine the optimal rotation and translation to bring the two data sets into registration.

The first invariant property that is exploited is each particle's nearest neighbors. To be precise, by nearest neighbor, we mean the particle with the smallest $\ell^{2}$-norm (Euclidean distance) between the centroid of the particles in two (or more) dimensions. A particle's nearest neighbors do not change regardless of rotation, translation, or even scaling of the entire data set. The nearest neighbors of a particle are unique and can be ordered by $\ell^{2}$-norm. If we examine the two nearest neighbors of each particle, these neighbors define a unique triangle.

Because the particle search process is not $100 \%$ efficient, we cannot assume that the measured data sets always accurately reflect the true nearest neighbors for each particle. Some particles will be missed meaning that some nearest-neighbor assignments will be incorrect. However, so long as a fraction of the nearest-neighbor groupings accurately reflects the

[^1]underlying reality, it should be possible to match up these nearest-neighbor groupings. The algorithm only requires a handful of matching groupings to work.

The second invariant property is the lengths of the edges of the triangle defined by the nearest neighbors. The length of the edges of a triangle defined (1) from the particle to the first nearest neighbor; (2) from the first nearest neighbor to the second nearest neighbor; and (3) from the particle to the second nearest neighbor is invariant under rotation and translation. In two dimensions, these three points are defined by two coordinates giving six degrees-of-freedom. We reduce the six degrees-of-freedom by three when we insist on rotational (one degree) and translational (two degrees) invariance, leaving a remainder of three degrees-of-freedom. The triangle's three edge lengths fully capture these three degrees-of-freedom. It is worth noting that the length of the edges is not rotationally invariant if the scale of the axes is not equal.

The third invariant that we exploit is the invariance of the center of mass. Once we have identified a grouping of corresponding particles between data sets, the global center of mass of the corresponding particles in each data set is invariant to rotation. The difference in the centers of mass can be used to determine the offset between the groupings regardless of relative rotation.

The fourth invariant property is the distance between groupings. This is to say that a pair of triangles of particles whose center-of-mass are separated by a distance $d$ in one particle data set will be separated by a distance $d$ in a secondeven if the second data set has been rotated and translated.

Finally, we expect the relative angle between paired groupings between data sets to be translationally invariant. This is the rotation angle between data sets.

This method is most similar to the class of graph matching algorithms (see, e.g., Zheng \& Doermann, 2006). Graph matching algorithms frame the problem in the language of graph theory in terms of vertices and edges. Similar to this algorithm, properties of the vertices and edges are used to match particles in one data set with the other. However, in our literature searches, we have not discovered another algorithm quite like this one.

## Implementation

Each particle data set is analyzed independently using the KDTree algorithm (Arnaldi et al., 1987) to determine its two nearest neighbors. The KDTree algorithm is frequently used in ray-tracing and 3D graphics applications. We use the Julia implementation provided by NearestNeighbors.jl (Carlsson et al., 2021). This algorithm sequentially bifurcates each coordinate axis to allow the identification of nearest neighbors using $\mathcal{O}(N \log N)$ operations rather than $\mathcal{O}\left(N^{2}\right)$ operations as the brute-force algorithm would. The KDTree is used within a data set to identify each particle's nearest two neighbors, which in turn identifies a unique triangle. The length of the three edges between the particles is computed and stored in a structure addressed by the index of the seed particle.

A second KDTree is constructed. This time the search is not performed in real space but in triangle edge length space. The KDTree is constructed from the edge lengths defined by the nearest neighbor triplets constructed in the previous step. The KDTree is evaluated against the other data set to discover the best matching pairs of triangles between measured data sets. Because the edge lengths are rotationally invariant, they

```
align(
    ps1::Vector{<:StaticVector{2,T}}, # Data set 1
    ps2::Vector{<:StaticVector{2,T}}; # Data set 2
    tol=0.001, # Approximate uncertainty in ps1 & ps2 data
    finealign=true # "true` performs a fine alignment
    where { T <: AbstractFloat }
correspondences
    ps1::AbstractVector{<:StaticVector{2, T}}, # Data set 1
    ps2::AbstractVector{<:StaticVector{2, T}}; # Data set 2
    tol=0.01, # Maximum distance between particles
    invert=false # 'true` to return non-correspondences
) where { T<: AbstractFloat }
identify(
    pss::AbstractVector{<:AbstractVector{<:StaticVector{2,T}}};
    tol=0.001,
    ctol=0.01
)::DataFrame where { T <: AbstractFloat}
```

Fig. 1. The function signatures for the align (. . ) , correspondences (. . ) , and identify (. . .) functions in the NeXLParticle library that implement the functionality described in this paper.


Fig. 2. The seven raw particle position data sets. The data sets are offset and translated relative to one another. You will also notice that there are no particles measured below the -30 mm line in the $Y$-dimension because of stage travel limitations.
can be matched between data sets and thus used to identify particles that are likely to represent the same particles in the two data sets. However, because of uncertainties in the positional measurements of each particle, there is no guarantee that all the triangles are good matches.
It should be noted that there is one ambiguity in the edge length comparison. Mirror symmetric triangles will produce equivalent edge lengths. This ambiguity can be eliminated using the determinant of the edge vectors to introduce a sign to distinguish clockwise and counter-clockwise aligned triangles.

Second- and third-level filters are used to further eliminate inferior matches. When we examine pairs of triangles between samples, we expect two things to be invariant. First, the distance between the centers-of-mass of each triangle should match between data sets. That is to say, the triangles are separated by some distance in the ground truth and should also be similarly separated in the measured data sets. Thus, we filter out pairs of triangles from both data sets for which the distance between each pair of groupings' center-of-mass does not match sufficiently well.

A third filter is applied by creating a histogram of angles defined by the vectors between the triangle pairs in data set 1 and data set 2 . The rotation angle should be consistent for equivalent triangle pairs between data sets. The angular bin with the most counts is identified as the rough estimate of the rotation angle between data sets. This estimate is further refined by averaging the angle for triangle pairs with angles close to the estimate. This angle is the algorithm's best estimate of the angle of rotation between the data sets.

Finally, we take all the triangle pairs which matched the three filters and compute the center-of-mass of all the triangles in each data set. The difference in center-of-mass equals the offset between the particle data sets. Thus, we have determined both the rotation and translational offset necessary to best register the two particle data sets.

We implemented the algorithm in the source file align. $j 1$ in the Julia library NeXLParticle (https://github.com/usnistgov/ NeXLParticle.jl) that is part of the NeXL collection of X-ray microanalysis libraries. Signatures and documentation for the functions align(...) and correspondences (...) are shown in Figure 1.


Fig. 3. The seven data sets registered using the algorithm described in this paper. In the plot, each subsequent data set has been offset by 0.1 mm in both the $X$ and $Y$ directions from the previous to help to visualize the overlap. As a result, corresponding particles in each data set appear to form a diagonal line.


Fig. 4. The histogram of distances when particle position data set $r 1$ is compared with data set $r 2$. While the majority of nearest-neighbor distances are less than $10 \mu \mathrm{~m}$, there are a number that are much larger. The larger values typically represent particles that were only measured in one data set or the other. The histogram plots for the remaining data sets look very similar. This suggests an excellent alignment of particles whose position is measured to a few micrometer accuracy in each dimension regardless of translation and rotation.

The code returns the result as two affine maps, where an affine map is a mathematical object that combines a linear transformation with a translation. The first affine map translates data set 1 to place its center-of-mass at the origin. The second affine map centers data set 2 at the origin and rotates the register individual particles with a particle in data set 1 . It is not possible to align the data sets with the "ground truth" as there is no way to define a canonical orientation or translation.

Overall, the algorithm has been demonstrated to align data sets consisting of approximately 100,000 particles in a fraction of a second.

This alignment is sufficiently accurate for many purposes. However, it can be further refined using a nonlinear optimization algorithm to minimize the distance between corresponding particles on data set 1 and data set 2 in a manner similar to Fitzgibbon (2003) used a Jacobian-informed MarquardtLevenberg optimizer. This is implemented using the finealign= true optional argument to the align (...) function.

Once aligned, two particle data sets can be compared based on the transformed coordinates. Two particles, one in each data set, are labeled as correspondences when the linear
distance between particles is less than a threshold. The appropriate threshold is determined by the positional uncertainty in each position measurement. The positional uncertainty is due to stage positioning inaccuracies, orthogonality imperfections between the stage axes, image magnification calibration inaccuracies, image orthogonality inaccuracies, and potentially other sources. While it is difficult to estimate a priori, the uncertainty can be readily extracted by constructing a histogram of the distances between nearest-neighbor particles in the aligned coordinate system.

The nearest-neighbor algorithm is used again to identify correspondences. A KDTree is constructed from one particle data set, and the closest particle meeting the distance threshold is identified as the corresponding particle in the other data set. The result is two equal length arrays of particle indices, one for each data set. The $i$ th index in each array corresponds to the $i$ th pair of corresponding particles. It is thus possible to identify particles have been measured twice, once in each analysis, and particles were only measured in one or the other analysis.

The correspondences (...) method has an argument invert which, when true, will produce the pair of indices of all particles for which there are no correspondences-they were only measured on one data set.

Finally, it is often useful to be able to track particle correspondences across two or more analyses. The function identify (...) takes a list of lists of particle coordinate pairs. It aligns the second through last list of particle coordinate pairs with the first list. It then uses correspondences to track particles across all the data sets. It returns a DataFrame with a row for each unique physical particle discovered in any of the analyses. Each row has a column for each data set which either contains the index of the particle in that data set or the value missing if that particle was not measured in that analysis.

## Validation

EGOS, europium-doped gadolinium oxysulfide $\left(\mathrm{Gd}_{2} \mathrm{O}_{2} \mathrm{~S}: \mathrm{Eu}\right)$, are a type of particle that can be imaged in a scanning electron


Fig. 5. Corresponding particles between data sets $r 1$ and $r 2$. The particle coordinates of the $r 2$ data set have been offset by 0.1 mm on each axis to allow the visualization of overlapping particles.


Fig. 6. Pairs of corresponding particles picked at random from $r 1$ (top row) and $r 2$ (bottom row). While the particles are translated, rotated, and not necessarily the most elegant images, they can be seen to represent the same particles in each data set.
microscope (SEM) and, when illuminated to induce fluorescence, readily identified in an optical microscope. EGOS particles are durable and immune to many common solvents. As such, they represent a useful proxy material for studying the efficiency of particle recovery from surfaces. The EGOS particles represent a challenging sample because they vary in dimensions from approximately $2 \mu \mathrm{~m}$ down to hundreds of nanometers. We expect to detect the larger particles with close to $100 \%$ efficiency. However, the smaller particles become more difficult to detect consistently using typical SEM-base automated particle analysis systems. The step size between pixels becomes comparable to the particle size, and the decrease in mass thickness of the particle causes the backscatter contrast and the X-ray signal to diminish. As a result, particles are not detected when the beam misses the particle, when the beam strikes the particle but the contrast is too low to trigger the threshold, and when the X-ray signal contains too few of the characteristic Gd and S X-rays. Project constraints mean that the sample area is necessarily large, the full area of a 25.4 mm diameter silicon wafer. The mean atomic number of the silicon substrate is relatively high compared to an alternative like a carbon substrate. As a result, the analysis time is long and we cannot afford to spend more time per sample to meet our throughput goals. Ultimately, the project goals require an accurate estimate of the number of EGOS particles on the substrate.

The sample in this study was prepared using a Collison nebulizer to disperse particles on a 25.4 mm silicon wafer. The sample was analyzed using the SEMantics extension to DTSA-II to automate particle data acquisition on a

TESCAN MIRA3 scanning electron microscope at 20 keV or 15 keV and 1 nA . The EGOS particles were distinguished from surface contamination using the combined energy dispersive X-ray signal from three Pulsetor silicon drift detectors. The spectra were quantified using measured standards, and the normalized k -ratios were evaluated using a thresholdbased rule to identify EGOS. Only the EGOS particles were included in the subsequent data analysis. The analysis area was defined by three points on the perimeter of the wafer that defines a circular region. The interior of the circular region was tiled with $256 \mu \mathrm{~m} \times 256 \mu \mathrm{~m}$ fields, which were scanned at $4 \mu \mathrm{~s} /$ pixel and a dimension of 1024 pixels $\times 1024$ pixels. The tiling produced approximately 7,200 tiles/analysis over an area of $490 \mathrm{~mm}^{2}$. One final analysis was performed using $128 \mu \mathrm{~m} \times$ $128 \mu \mathrm{~m}$ fields to improve sensitivity for the smallest particles by decreasing the steps size between beam raster points. The backscatter detector signal and a threshold-based algorithm were used to discriminate particles. The backscatter detector gain and contrast were set such that Si produced a signal of 16 on an 8 -bit digitizer and Ge produced a signal of 225.

The sample was analyzed seven times at different translations and rotations. The first five at 20.0 keV , the last two analyses at 15.0 keV , and the final analysis was tiled with $128 \mu \mathrm{~m} \times 128 \mu \mathrm{~m}$ fields. The total counts of EGOS particles found in each analysis were $1,771,1,799,1,798,1,850$, $1,823,1,740$, and 2,035 , respectively. This is interesting in-and-of-itself. In particular, the final count of 2,035 when the magnification was increased by a factor of 2 shows that particles are being missed in the lower magnification analyses. Figures 2 and 3 show the raw particle position data as collected and the particle position data after registration. The data sets have been incrementally offset in Figure 3 to allow the overlapping particles to be visualized.

Once the data sets have been registered, it is now possible to identify corresponding particles between the data sets. As each particle is discovered during an analysis, it is assigned a sequential index to identify it within the data set. Once data sets have been aligned, corresponding particles are likely to


Fig. 7. Particles in data sets $r 1$ and $r 2$ that do not have corresponding particles in the other data set. Three distinct classes of noncorresponding particles are readily identified. The first class represents those particles below the -30 mm extent of the stage motions (below - -30 mm in Y ). The second class represents those particles around the perimeter of the sample which may have been excluded by the field tiling, which is designed to maximally fill the circular analysis area without extending outside of the area. This means that subtle differences in the positioning of the three points that define the perimeter of the circular analysis area can lead to regions around the perimeter that are unanalyzed. Finally, the final class represents those particles in the interior for which other explanations are necessary. The particle may have been simply missed for one reason or another, or the particle may have moved between analyses.

Table 1. Fourteen Examples of Tracking Individual Particles Selected from the 2,573 Distinct Particles Measured in the Seven Repeat Analyses Labeled $r 1$ to $r 7$.

| $r 1$ | $r 2$ | $r 3$ | $r 4$ | $r 5$ | $r 6$ | $r 7$ | Reps | $X_{\text {tr }}$ | $Y_{\text {tr }}$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  |  |  |  |  | 351 | 1 | 3.872 | -23.891 |
|  |  |  |  |  |  | 459 | 1 | 0.385 | -26.448 |
| 391 |  |  |  |  |  | 850 | 2 | -3.775 | -25.513 |
| 411 |  |  |  |  |  | 676 | 2 | -1.310 | -25.134 |
| 130 |  |  |  |  | 358 | 366 | 3 | -0.246 | -28.538 |
| 1,114 |  |  |  | 959 |  | 1,446 | 3 | -3.642 | -17.711 |
|  | 1,146 | 1,124 | 1,495 |  |  | 303 | 4 | 4.995 | -23.345 |
| 1,494 | 264 |  | 486 |  | 1,203 |  | 4 | 0.512 | -13.339 |
| 280 | 1,406 | 1,394 |  | 909 |  | 78 | 5 | 6.149 | -26.421 |
|  | 363 |  | 786 | 252 | 843 | 951 | 5 | 4.145 | -14.838 |
|  | 1,561 | 1,552 | 1,509 | 1,369 | 409 | 417 | 6 | -0.094 | -27.483 |
| 1,019 | 696 | 703 | 1,075 |  | 655 | 716 | 6 | 3.888 | -18.647 |
| 321 | 1,533 | 1,522 | 1,009 | 1,730 | 1,019 | 1,171 | 7 | -7.836 | -26.223 |
| 1,369 | 340 | 357 | 812 | 187 | 819 | 904 | 7 | 4.864 | -14.587 |

The table contains two examples of each number of repeat measurements from once through seven times selected at random from the distinct particles measured in the seven analyses. The number in the $r X$ column represents the index of the particle in that data set. The $X_{\mathrm{tr}}$ and $Y_{\mathrm{tr}}$ columns represent the mean of the translated coordinates from each data set in which the particle was measured.
have very similar transformed coordinates. The degree to which this is true can be evaluated by plotting the histogram of the $\ell^{2}$-norm between inter-data set nearest neighbors. Well-aligned data sets will be dominated by many correspondences with a small $\ell^{2}$-norm. Figure 4 shows the histogram of distances between nearest neighbor particles when data set $r 1$ is compared to data set $r 2$.
Based on the results presented in Figure 4, a distance threshold of 0.02 mm was selected to determine corresponding particles between particle data sets. Figure 5 shows the corresponding particles between data sets $r 1$ and $r 2$. To demonstrate that the correspondences do in fact represent the same particle measured in different analyses, we examined a
random selection of particle image pairs. Figure 6 shows corresponding pairs of particles picked at random from all corresponding pairs in $r 1$ and $r 2$. The images may be rotated but clearly represent the same particles measured in different analyses. Figure 7 shows the particles in data sets $r 1$ and $r 2$ that do not have nearest neighbors in the other data set within the threshold of 0.02 mm .
Finally, we would like to be able to track the same particle through multiple analyses. Identifying correspondences between data sets makes it possible to track individual particles through all seven data sets. Table 1 shows a small subset of the 2,573 distinct particles identified in the seven analyses. Some particles were measured many times during the seven analyses


Fig. 8. The left plot displays the first analysis in which a particle is measured. Most of the particles were seen in the first two analyses, but there were a number of particles that were not seen until the sixth, in which the beam energy was reduced to 15 keV to reduce the excitation volume, and the seventh, in which the beam energy was 15 keV and the magnification was increased by a factor of 2 . The plot on the right shows the number of times a particle was measured.


Fig. 9. This plot shows a histogram of the particle size for particles that were first measured in $r 1, r 3, r 5$, and $r 7$. It shows that it is mostly the smaller particles that were missed during early analyses and that a lot of small particles were not picked up until the magnification was increased by a factor of 2 in $r 7$.
and some were measured as few as once. Figure 8 gives a better idea of in which analysis a particle was first seen and how many times the particle was measured. Figure 9 shows that the size of the particle is a strong predictor of being missed. Many small particles were missed by the first six analyses and only picked up by the seventh, in which the magnification was increased by a factor of 2 .

## Conclusion

In this paper, we have reported an algorithm that is well suited to aligning particle data sets such as are collected in electron and optical microscopy. The algorithm depends upon the fact that the measured data sets reflect an underlying distribution which permits one-to-one correspondences between many of the data points in one data set with the other. This one-to-one correspondence does not need to be initially
known but is determined by the algorithm. This is distinct from the majority of the algorithms in the literature in which the point clouds are determined from arbitrarily selected points on a surface.

While this algorithm has been observed to work very well in our laboratory, there are many subtle reasons it may fail. First, the ideal sample will have small uncertainty relative to the distance to the nearest neighbor. This makes it likely that the particle triplets will be matched correctly between samples and less likely that a particle triplet will be incorrectly matched. Samples with more particles are typically easier to align than samples with fewer because there are more opportunities for matched pairs. This is true even when there may be false matches because all true matches must undergo the same rotation. Thus matched triplets with different rotations can be rejected as false matches. Once the rotation filter has been applied, a second filter ensures that all the translations are similar. These two filters are very efficient at rejecting false matches which may occur more frequently in large particle data sets. Second, the lengths of the two stage axes must be equal. When they are not equal, rotations also introduce changes in the nearest neighbor distances hindering the match process. Third, particles must be measured with sufficient efficiency that the two nearest neighbors frequently represent the same particles. It is necessary that a particle and its two nearest neighbors are measured on both samples. If the likelihood of detection is $P$, then the likelihood that the two nearest neighbors are detected is $P^{2}$. Fortunately, only a relatively small number of true matches need to be found for the algorithm to work. Fourth, since the measured particle position is a function of both stage and image calibrated coordinates, the stage and image must have similar calibrated scales. There are often two subtly incommensurate length scales in microscopy-the scale defined by the optics and the scale defined by the stage. Nominally, these are similar and can be calibrated to make them equivalent. However, this is particularly challenging
with electron microscopy, where the magnification scale for every beam energy and working distance must be computed using models of the electron optics. Fifth, in electron microscopy, there is the additional complication of uncompensated image field rotation caused by the precession of electrons in the magnetic field of the objective lens. This can lead to image fields that are not perfectly registered with the stage axes.

Commensurate scaling is particularly important when the data sets are measured on different instruments. Since nearestneighbor matching is a local process (over short distances), this is likely to be relatively unaffected by differences in axis scaling of a few percent between instruments. However, since the distance between particle triplets on opposites sides of the sample may be much larger, the absolute registration accuracy will suffer even when the scales between instruments only differ by a percent or two.

Experience with the algorithm suggests that it is highly reliable for well dispersed collections of hundreds to hundreds-of-thousands particles with measurement accuracy of 1-part-in-100 of the mean nearest-neighbor distance. The degradation of the algorithm with adverse measurement conditions has not been extensively studied. To evaluate whether it is likely to work on your data, it may be efficient to simulate data sets with particle loadings similar to your samples measured on an instrument with realistic limitations. Apply the algorithm to an ensemble of these simulated data sets to estimate the algorithm's reliability.

While the algorithm has been implemented for twodimensional data sets collected on instruments with commensurate equi-axis scales, a few obvious extensions suggest themselves.

1. Different scales between data sets. It should be possible to extend this algorithm to measurements of the same particle data set on instruments with different unknown scales. This would reduce the information content in the data by one additional degree of freedom. Instead of using the three length of edges as a metric, two independent ratios of lengths of edges or the angles between edges could be used.
2. It could be useful in some cases to use more than two nearest neighbors. Two nearest neighbors are defined by three coordinates. Three nearest neighbors could be defined by four coordinates defining three triangles.
3. It is possible to extend the algorithm to three or higher dimensions. Three dimensions would require extracting a three-dimensional offset and two rotation angles, but
each nearest neighbor point would carry an additional coordinate's worth of data.

Finally, our example data sets demonstrate that counting particles, particularly small particles near the detection threshold, is difficult. Small changes in focus and random fluctuations mean that the detection probability of small particles drops precipitously after a certain parameter-dependent size. Producing reproducible particle counts near this size is very challenging.

## Acknowledgments

This work was undertaken as part of the authors' official duties as employees of the National Institute of Standards and Technology, an agency of the United States Department of Commerce.

## Conflict of interest

The authors certify that they have no competing interests that could impact any aspect of this work.

## References

Arnaldi B, Priol T \& Bouatouch K (1987). A new space subdivision method for ray tracing CSG modelled scenes. Vis Comput 3, 98-108.
Besl PJ \& McKay ND (1992). Method for registration of 3-D shapes. In Sensor Fusion IV: Control Paradigms and Data Structures, Schenker PS (Ed.), vol. 1611, pp. 586-606. International Society for Optics and Photonics, SPIE. Available at https://doi.org/10.1117/12.57955
Bezanson J, Edelman A, Karpinski S \& Shah VB (2017). Julia: A fresh approach to numerical computing. SIAM Rev 59, 65-98.
Carlsson K, Karrasch D, Bauer N, Kelman T, Schmerling E, Hoffimann J, Visser M, San-Jose P, Christie J, Ferris A, Anthony Blaom P, Pasquier B, Foster C, Saba E, Goretkin G, Orson I, Samuel O, Choudhury S \& Nagy T (2021). KristofferC/NearestNeighbors.jl: v0.4.9. Available at https://doi.org/10.5281/zenodo. 4943232
Fitzgibbon AW (2003). Robust registration of 2D and 3D point sets. Image Vis Comput 21, 1145-1153.
Maiseli B, Gu Y \& Gao H (2017). Recent developments and trends in point set registration methods. J Vis Commun Image Represent 46, 95-106.
Moré JJ (1978). The Levenberg-Marquardt algorithm: Implementation and theory. Numerical Analysis, pp. 105-116. Berlin: Springer.
Zheng Y \& Doermann D (2006). Robust point matching for nonrigid shapes by preserving local neighborhood structures. IEEE Trans Pattern Anal Mach Intell 28, 643-649.

## DIATOME Diamond KNives

the incomparable Diamond Knife for all fields of research...

> Please visit our website for our complete range of Diamond Knives: www.diatomeknives.com

## ultra Maxi

Similar to our ultra $35^{\circ} 4.0 \mathrm{~mm}$ but with a larger boat. Applications include soft industrial samples such as metals and polymers, hard and brittle samples such as semiconductors, superconducting oxides, nanocrystalline ceramics.


Rat muscle (Quadriceps) x 23 '000 Werner Graber, Anatomisches Institut, Bern


Eye of A. peroni: part of a sequence of semithin sections. Michael J.F. Blumer, Institut für Zoologie, Universität Wien Reprinted from: Ribbons of semithin sections an advanced method with a new type of diamond knife. Journal of Neuroscience Methods 120 (2002 11-16), with permission from Elsevier. See the complete series on diatomeknives.com

## histo Jumbo

Perfect for Immuno-histo-chemistry
For 3D reconstruction it is imperative not to lose a single section (Ref. Blumer). The large Jumbo boat as well as the adhesive (Pattex compact by Henkel) applied to the side of the sample block increase the distinct advantages. They allow:

- Easy production of section ribbons (0.5-2 $\mu \mathrm{m}$ )
- No section loss
- No folding
- The same orientation of all sections
- Easy collection of section ribbons
- Multiple ribbons on one glass slide


## DIATOME U.S.

Nondecalcified rat bone. Scale: $35 \mathrm{~mm}=100 \mu \mathrm{~m}$. Daniel Studer, Anatomisches Institut, Bern.

The knife is designed for the sectioning of hard and soft biological and industrial materials, non embedded or embedded in methacrylate or epoxy resins. The histo knife may be used on all ultramicrotomes and microtomes with a retraction of the specimen in the return phase.
Advantages compared to glass knives:

- Perfect sections, free of scores or compression.
- Serial sections without knife change.
- Thinner sections.



[^0]:    Received: June 14, 2022. Revised: September 12, 2022. Accepted: October 24, 2022
    Published by Oxford University Press on behalf of the Microscopy Society of America 2023.
    This work is written by (a) US Government employee(s) and is in the public domain in the US.

[^1]:    1 Disclaimer: Any mention of commercial products is for information purposes only; it does not imply recommendation or endorsement by NIST.

