Joint registration of multiple point clouds for fast particle fusion in localization microscopy

Wang, Wenxiu; Heydarian, Hamidreza; Huijben, Teun A.P.M.; Stallinga, Sjoerd; Rieger, Bernd

DOI
10.1093/bioinformatics/btac320

Publication date
2022

Document Version
Final published version

Published in
Bioinformatics

Bioinformatics, 38(12), 2022, 3281–3287
https://doi.org/10.1093/bioinformatics/btac320
Advance Access Publication Date: 13 May 2022
Original Paper

Bioimage informatics

Joint registration of multiple point clouds for fast particle fusion in localization microscopy

Wenxiu Wang , Hamidreza Heydarian, Teun A.P.M. Huijben, Sjoerd Stallinga* and Bernd Rieger*

Department of Imaging Physics, Delft University of Technology, Delft 2628CJ, The Netherlands

*To whom correspondence should be addressed.

Associate Editor: Alfonso Valencia

Received on September 15, 2021; revised on February 22, 2022; editorial decision on May 2, 2022; accepted on May 9, 2022

Abstract

Summary: We present a fast particle fusion method for particles imaged with single-molecule localization microscopy. The state-of-the-art approach based on all-to-all registration has proven to work well but its computational cost scales unfavorably with the number of particles N, namely as N^2. Our method overcomes this problem and achieves a linear scaling of computational cost with N by making use of the Joint Registration of Multiple Point Clouds (JRMPC) method. Straightforward application of JRMPC fails as mostly locally optimal solutions are found. These usually contain several overlapping clusters that each consist of well-aligned particles, but that have different poses. We solve this issue by repeated runs of JRMPC for different initial conditions, followed by a classification step to identify the clusters, and a connection step to link the different clusters obtained for different initializations. In this way a single well-aligned structure is obtained containing the majority of the particles.

Results: We achieve reconstructions of experimental DNA-origami datasets consisting of close to 400 particles within only 10 min on a CPU, with an image resolution of 3.2 nm. In addition, we show artifact-free reconstructions of symmetric structures without making any use of the symmetry. We also demonstrate that the method works well for poor data with a low density of labeling and for 3D data.

Availability and implementation: The code is available for download from https://github.com/wexw/Joint-Registration-of-Multiple-Point-Clouds-for-Fast-Particle-Fusion-in-Localization-Microscopy.

Contact: s.stallinga@tudelft.nl or b.rieger@tudelft.nl

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The diffraction of light limits the resolution of conventional microscopy to about 200 nm. Several super-resolution microscopy techniques enable 'diffraction unlimited' resolution (Hell, 2009; Klein *et al.*, 2014; Vicidomini *et al.*, 2018). Single-molecule localization microscopy is a widely used member of the family of super-resolution techniques, and obtains super-resolved images by localizing single fluorescent emitters. The resolution of these super-resolved images is not infinite, but in practice restricted to about 20 nm due to the incomplete fluorescent labeling and a limited number of collected photons per localization event (Nieuwenhuizen *et al.*, 2013). In recent years, significant improvements have been made to increase the photon count per localization (Metzger *et al.*, 2016). Increasing the density of labeling (DOL) using biochemical means is difficult, where DOL values of around 50% are typically achieved. In addition, a high local DOL can lead to an increased rate of mislocalizations (Fox-Roberts *et al.*, 2017) which is detrimental for the quality of the imaging process. If the sample includes many chemically identical biocomplexes (called particles in the following), the limitation imposed by a low DOL can be lifted by fusion of all these particles into one single reconstruction, the so-called super-particle, leading to a much better resolution and signal-to-noise ratio (Löschberger *et al.*, 2012; Szymborska *et al.*, 2013). This approach by particle fusion, of course, ignores potential heterogeneity in the underlying biology within the collection of particles. Template-driven particle fusion methods have been used (Broeken *et al.*, 2015; Gray *et al.*, 2016; Löschberger *et al.*, 2012; Szymborska *et al.*, 2013) but have a substantial risk of resulting in a biased reconstructed structure. Heydarian *et al.* (2018b, 2021a) proposed a template-free particle fusion method based on an all-to-all registration (all-to-all method in short), which is robust against under-labeling and misregistration. The all-to-all method has proven to work well and produces reconstruction resolutions down to a few nanometers. Despite this success, computational times of around a

day for a number of particles N exceeding about 1000 are not uncommon and are only feasible with the use of GPU acceleration. The root cause lies within the unfavorable scaling of computational cost with N^2, because each particle is registered to all other particles, resulting in $N(N-1)/2$ registration pairs. The all-to-all method has another drawback, the so-called 'hot-spot' problem. For symmetric structures, random variations in the localization data with binding site are amplified by the pair-wise optimal registration process. Heydarian *et al.* solved this problem by first detecting the present symmetry and then imposing it on the data in a post-processing step. Thus, a particle fusion algorithm that is fast and which avoids the hot-spot artifact is desired.

An alternative to the all-to-all method is based on the Joint Registration of Multiple Point Clouds (JRMPC) method (Evangelidis and Horaud, 2017). In the JRMPC method, particles are iteratively rotated and translated to fit to a Gaussian Mixtures Model (GMM), which is updated itself in each iteration round. The key advantage of the JRMPC method is that the computational complexity scales linearly with the number of particles N, which makes it inherently faster than the all-to-all method if N grows large. In addition, hot-spot artifacts in symmetric structures are avoided without imposing (*a priori*) symmetry information, because the joint registration treats each particle equally. There are, however, major drawbacks to the JRMPC method. First, the outcome of the JRMPC turns out to be highly susceptible to the initialization of the GMM (number of Gaussians, center positions and widths). Different initial settings of the GMM parameters lead to different sets of final estimated particle rotations and translations. Second, the final outcome usually consists of several clusters, where the particles within the clusters are well-registered, but where the clusters have different poses. We attribute these issues with robustness of the algorithm to trapping in local optima of the iterative optimization (outlined in Section 2.1 in detail).

The goal of the work presented in this article is to overcome the robustness problems of the JRMPC method while maintaining the inherent speed advantage. To this end, we propose a processing pipeline in which we combine JRMPC registration outcomes obtained with different GMM initializations using cluster analysis tools. The cluster analysis uses our recent unsupervised classification framework (Huijben *et al.*, 2021), which is based on the Bhattacharya distance metric (Broeken *et al.*, 2015) together with multidimensional scaling (MDS) (Mead, 1992) and k-means clustering (Cheng, 1995; Jain *et al.*, 1999). The process of JRMPC and classification is repeated several times for different GMM initializations. Pairs of clusters from different initializations may share particles. The relative poses of such particles in different clusters is used in a final step to combine the different clusters into a single well-aligned structure.

2 Materials and methods

Our proposed algorithm has three main steps, illustrated in Figure 1. The steps are (1) alignment of particles with JRMPC using multiple initializations, (2) classification of JRMPC registered particles into clusters and (3) connection of the identified clusters into a single final reconstruction.

The input data are a union of particles $\mathbf{A} = \{\mathbf{A}_j\}_{j=1}^{N}$, with N the number of particles. Each particle is characterized by a set of localization coordinates $\mathbf{V}_j$ and attendant localization uncertainties Δ_j as $\mathbf{A}_j = \{\mathbf{V}_j; \Delta_j\}$. The coordinates of particle j represent M_j localizations:

$$\mathbf{V}_j = [\mathbf{v}_{j1} \dots \mathbf{v}_{ji} \dots \mathbf{v}_{jM_j}] \in \mathbb{R}^{d \times M_j},$$

where the $\mathbf{v}_{ji}$ are vectors with elements equal to the d coordinates of the i-th localization in particle j. Depending on the data, the dimensionality d can be 2 or 3. In general, the localization uncertainties of the M_j localization events in particle j are:

$$\Delta_j = [\Sigma_{j1}, \dots \Sigma_{ji} \dots \Sigma_{jM_j}] \in \mathbb{R}^{d \times d \times M_j},$$

where the Σ_{ji} are $d \times d$ matrices equal to the covariance matrices of the i-th localization in particle j. Often a more simple description of the localization uncertainty is possible. For 2D data for example, the uncertainties are isotropic, and Δ_j can be written as:

$$\Delta_j = [\delta_{j1}, \dots \delta_{ji} \dots \delta_{jM_j}],$$

where the δ_{ji} are now scalar values that represent the localization uncertainty in the xy plane for the i-th localization in particle j. For most 3D data, Δ_j is represented as:

$$\Delta_j = [\delta_{j1}, \tau_{j1}; \dots, \delta_{ji}, \tau_{ji}; \dots \delta_{jM_j}, \tau_{jM_j}],$$

where now τ_{ji} is the localization uncertainty along the z-axis for the i-th localization in particle j. This axial localization uncertainty is typically larger than the uncertainty in the xy plane (Rieger and Stallinga, 2014).

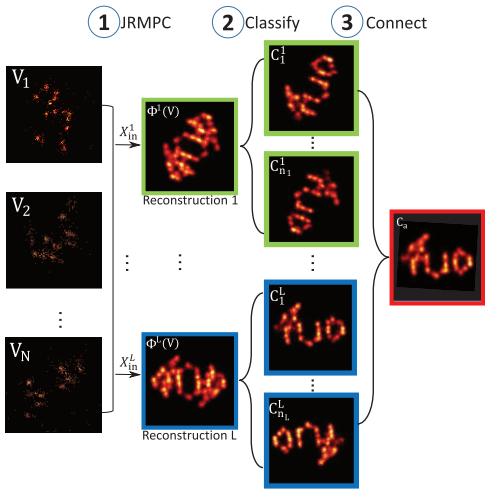

Fig. 1. The three main steps of the proposed particle fusion algorithm. Step 1: Use JRMPC (Evangelidis and Horaud, 2017) to initially align N input particles $\mathbf{V} = \{\mathbf{V}_j\}_{j=1}^{N}$ with L random initializations of the GMM $\{\mathbf{X}_{\text{in}}^{l}\}_{l=1}^{L}$ leading to L different reconstructions $\{\mathbf{\Phi}^{l}(V)\}_{l=1}^{L}$. Step 2: Apply the unsupervised classification method of Huijben *et al.* (2021) to classify each reconstruction $\Phi^{l}(V)$ into n_l clusters $\{C_n^l\}_{n=1}^{n_l}$ separating different overlapping poses in the reconstructed particles. Step 3: Connect particles from different clusters into the final super-particle reconstruction C_a, such that each input particle is present at most once

2.1 Alignment

The structure of the reconstruction is characterized in the JRMPC method by a GMM with parameters $\mathbf{G} = \{\mathbf{G}_k\}_{k=1}^{K}$, where each of the K Gaussians components $\mathbf{G}_k = [p_k, \boldsymbol{\mu}_k, \sigma_k]$ has a mixing coefficient (weight) p_k, a set of d coordinates $\boldsymbol{\mu}_k$ that represent the mean of the Gaussian, and a standard deviation σ_k (an isotropic covariance matrix $\sigma_k^2\mathbf{I}_d$ is taken). The GMM parameters have an initial setting $\mathbf{G}_{\text{in}}$, described in Section 3.2. The parameters that are updated during the iterative JRMPC algorithm are:

$$\mathbf{\Theta} = \left\{\{\mathbf{G}_k\}_{k=1}^{K}, \{\mathbf{R}_j, \mathbf{t}_j\}_{j=1}^{N}\right\}, \tag{1}$$

where $R_j \in \mathbb{R}^{d \times d}$ is the rotation applied to particle j and where $t_j \in \mathbb{R}^{d \times 1}$ is the translation applied to particle j. The coordinates of the reconstruction are then:

$$\mathbf{\Phi}(\mathbf{V}) = \{\mathbf{R}_j\mathbf{V}_j + \mathbf{t}_j\}_{j=1}^{N}, \tag{2}$$

which thus contains the coordinates of all localization events in all particles. It is noted that the localization uncertainties are not taken into account in the JRMPC method. Further details on the steps in each iteration round of the JRMPC are given in Supplementary Appendix A.

The outcome of the JRMPC depends on the choice of the initial GMM centers in $\mathbf{G}_{\text{in}}$. Our algorithm uses L differently initialized GMMs $\{\mathbf{G}_{\text{in}}^{l}\}_{l=1}^{L}$, leading to L different JRMPC alignments $\mathbf{\Phi}(\mathbf{V}) = \{\mathbf{\Phi}^{l}(\mathbf{V})\}_{l=1}^{L}$ of the same union of particles with coordinates $\mathbf{V}$.

2.2 Classification

The JRMPC algorithm can end up in a local optimum, resulting in multiple groups of particles (clusters) with different overlapping poses in the reconstruction. To separate these clusters, we use an unsupervised classification method recently proposed by our group (Huijben *et al.*, 2021). This method enables the analysis of structural heterogeneity in localization datasets arising from e.g. naturally occurring biological variations. Here, we use this pipeline to decompose the L different JRMPC outcomes into clusters of particles, where the particles within each cluster are well-aligned. First, we compute the normalized Bhattacharya cost function between every transformed particle $\Phi_a^l(\mathbf{V}_a)$ and every other transformed particle $\Phi_b^l(\mathbf{V}_b)$ within the JRMPC registration for each initialization $l = 1, 2, \ldots, L$. This one time computation gives an upper triangular matrix with $N(N-1)/2$ cost function values S. The normalized Bhattacharya cost is in general given by the sum over the M_a localizations of particle a and M_b localizations of particle b as:

$$S(a,b) = \frac{1}{M_a M_b} \sum_{q=1}^{M_a} \sum_{r=1}^{M_b} \frac{1}{\sqrt{\det \Omega_{qr}^{ab}}} \quad \exp\left(-\frac{1}{2} \delta\phi_{qr}^{ab\,T} \Omega_{qr}^{ab} \delta\phi_{qr}^{ab}\right). \quad (3)$$

Here $\delta\phi_{qr}^{ab} = \phi(\mathbf{v}_{aq}) - \phi(\mathbf{v}_{br})$ is the difference in transformed (rotated and translated) coordinates of localization q of particle a and localization r of particle b, and Ω_{qr}^{ab} is defined in terms of the uncertainty covariance matrices of the localizations as:

$$\Omega_{qr}^{ab} = \Sigma_{aq} (\Sigma_{aq} + \Sigma_{br})^{-1} \Sigma_{br}. \quad (4)$$

For example, for 2D data with isotropic localization uncertainties, this reduces to:

$$S(a,b) = \frac{1}{M_a M_b} \sum_{q=1}^{M_a} \sum_{r=1}^{M_b} \frac{1}{(\delta_{aq}^2 + \delta_{br}^2)} \exp\left(-\frac{(\phi(\mathbf{v}_{aq}) - \phi(\mathbf{v}_{br}))^2}{2(\delta_{aq}^2 + \delta_{br}^2)}\right). \quad (5)$$

The normalization of the cost function with the numbers of localizations per particle reduces the impact of the variations in these number, which makes it a better descriptor of the similarity between the structure of the particles. The next step is to transfer dissimilarity values:

$$D(a,b) = \max(S) - S(a,b) \quad (6)$$

to spatial coordinates in a multidimensional space suitable for classification using MDS (Mead, 1992). The transformed particles will then be partitioned into clusters by k-means clustering (Cheng, 1995; Jain *et al.*, 1999) in this multidimensional space. Parameter settings for the classification step are given in Section 3.2. This process is repeated for the JRMPC reconstructions $l = 1, 2, \ldots, L$ leading to $n = 1, 2, \ldots, n_l$ clusters that are denoted as C_n^l (see Fig. 1).

2.3 Connection

As we repeat the JRMPC reconstruction L times, pairs of clusters from different initializations may share different particles. Therefore, we need to combine the different clusters into a single well-aligned structure. In a first step, we discard clusters with less than ϑ particles. This threshold helps to filter out poorly aligned clusters as well as clusters with particles of poor quality, as these tend to accumulate in clusters with low number of particles.

Next, the cluster with the largest number of particles is selected as initial estimate of the super-particle reconstruction $\mathbf{C}_a$. This main cluster, $\mathbf{C}_m^{l_m}$, is used as the target for a pairwise comparison of clusters. A loop over all clusters $\mathbf{C}_n^l$ for $l \neq l_m$ is done, and clusters $\mathbf{C}_n^l$ and $\mathbf{C}_m^{l_m}$ are compared to check for particles that are in both clusters. If there exists at least one common particle c with coordinates $\mathbf{V}_c \in \mathbf{V}$ then the clusters $\mathbf{C}_n^l$ can be added to the super-particle reconstruction estimate $\mathbf{C}_a$ following:

Step 1: apply the inverse transformation of particle c in cluster $\mathbf{C}_n^l$ to transform all particles in the cluster $\mathbf{C}_n^l$ to the original position and pose of $\mathbf{V}_c$:

$$\mathbf{C}_n^l|_{\mathbf{V}_c} = \{\mathbf{R}_c^l\}^{-1} \mathbf{C}_n^l - \mathbf{t}_c^l. \quad (7)$$

Step 2: apply the transformation of particle c in the main cluster $\mathbf{C}_m^{l_m}$ to all particles in the cluster $\mathbf{C}_n^l|_{\mathbf{V}_c}$ to the position and orientation of $\mathbf{C}_m^{l_m}$:

$$\mathbf{C}_n^l|_{\mathbf{C}_m^{l_m}} = \mathbf{R}_c^{l_m} \mathbf{C}_n^l|_{\mathbf{V}_c} + \mathbf{t}_c^{l_m}. \quad (8)$$

Now that the cluster $\mathbf{C}_n^l$ is aligned with the pose of the main cluster $\mathbf{C}_m^{l_m}$ the particles of $\mathbf{C}_n^l$ can be added to the super-particle reconstruction estimate $\mathbf{C}_a$. In this way more and more particles accumulate in the final reconstruction, yielding the final outcome of our proposed algorithm.

Care must be exercised for two subtleties. First, it can happen that there is more than one common particle between the two clusters $\mathbf{C}_n^l$ and $\mathbf{C}_m^{l_m}$. Then, if there exists more than one common particles between two clusters, we will calculate all the common particles' translation matrices and rotation matrices from the cluster $\mathbf{C}_a^{l_1}$ to the cluster $\mathbf{C}_a^{l_1}$,

$$\mathbf{t}_c|_{\mathbf{C}_a^{l_1} \to \mathbf{C}_b^{l_2}} = \mathbf{t}_c^{l_2} - \mathbf{R}_c^{l_2} \{\mathbf{R}_c^{l_1}\}^{-1} \mathbf{t}_c^{l_1}, \quad (9)$$

$$\mathbf{R}_c|_{\mathbf{C}_a^{l_1} \to \mathbf{C}_b^{l_2}} = \mathbf{R}_c^{l_2} \{\mathbf{R}_c^{l_1}\}^{-1}, \quad (10)$$

then we compare all the $\mathbf{t}_c$ and $\mathbf{t}_c$ and use the common particle with rotation and translation matrix that are closest to the median of all translation and rotation matrices of all the common particles. Second, we must check if the particles of cluster $\mathbf{C}_n^l$ are not already in the reconstruction estimate $\mathbf{C}_a$. Only the unique particles that are not already contained in the reconstruction are added to $\mathbf{C}_a$. In Supplementary Appendix B, we give pseudo code for this connection pipeline.

3 Experiments

3.1 Experimental data

We applied our method to four different localization microscopy experiments described here:

DNA origami TUD-logo: We tested three different 2D TUD-logo DNA origami datasets (Heydarian *et al.*, 2018b) with *DOL* of 30%,

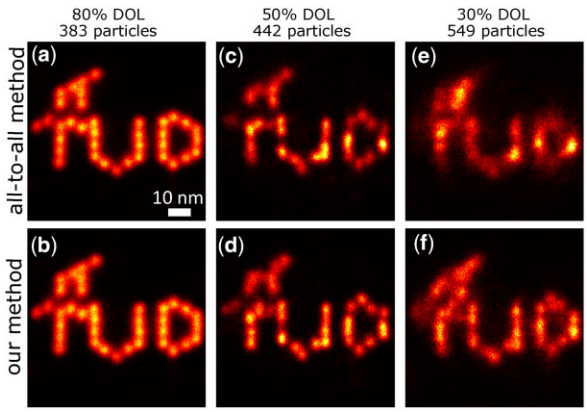

Fig. 2. Comparison of the particle fusion performance with our method and the all-to-all method on experimental 2D TUD-logo DNA origami particles. (**a**, **c**, **e**) Reconstruction by all-to-all registration (FRC resolution of 3.3, 3.5, 5.0 nm for 80%, 50% and 30% *DOL*, respectively). Computational time for (a) is about 2 h (GPU)). (**b**, **d**, **f**) Reconstruction by our method (FRC resolution of 3.2, 3.1, 3.3 nm for 80%, 50% and 30% *DOL*, respectively). Computational time for (b) is about 9.5 min (CPU). Scale bar of (a) applies to (b–f)

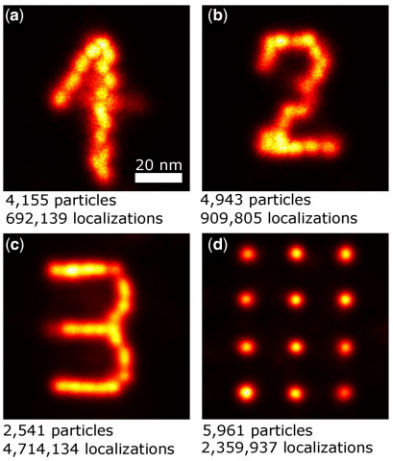

Fig. 3. Particle fusion speed for experimental 2D DNA-origami with a large number of particles. (**a**) Reconstruction of digit 1, computational time 1.1 h (CPU). (**b**) Reconstruction of digit 2, computational time 1.3 h (CPU). (**c**) Reconstruction of digit 3, computational time 48 min (CPU). (**d**) Reconstruction of 3×4 grid, computational time 4.8 h (CPU). The number of particles and localizations in each reconstruction are indicated below the figures. Scale bar of (a) applies all sub-images

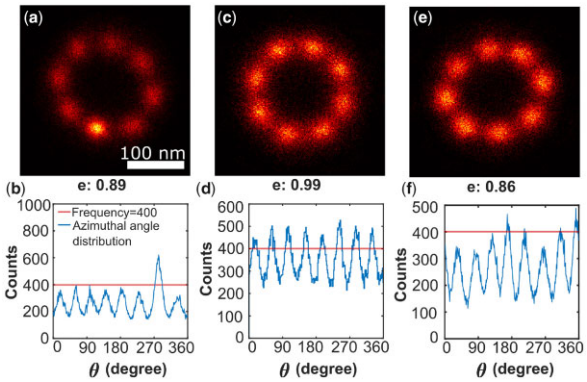

Fig. 4. Comparison of particle fusion performance between our method and the all-to-all method on 304 experimental 2D nuclear pore complex particles. (**a**) Reconstruction with the all-to-all method without prior knowledge. A 'hot-spot' is visible due the enhancement by pair-wise registration. Fitted ellipticities e to the reconstruction are shown below. (**b**, **d**, **f**) Histogram of the azimuthal angles of the localizations in (**a**, **c**, **e**), respectively; for comparison, a red line indicates 400 counts. (**c**) Reconstruction with the all-to-all method after explicitly imposing eight-fold symmetry. (**e**) Reconstruction with our method without prior knowledge. Even without imposing symmetry no hot-spot occurs. Scale bar of (a) applies to (c, e)

50% and 80%. We compared the results of the currently proposed method and the all-to-all method (Heydarian *et al.*, 2018b) in Figure 2. The data are available online (Heydarian *et al.*, 2018a).

2D nuclear pore complex: We further applied our method to 2D Nuclear Pore Complex (NPC) data which were previously described in Löschberger *et al.* (2012). In Figure 4, we show our reconstruction of NPCs together with the reconstruction of the all-to-all method (Heydarian *et al.*, 2018b) to compare the methods' capabilities in the reconstruction of symmetrical structures.

3D nuclear pore complex: We applied our algorithm to 3D NUP107 NPC data (Heydarian *et al.*, 2021a) acquired by two different localization microscopy techniques. The data are available online (Heydarian *et al.*, 2021c). The poses of the NPCs are experimentally constrained as they are all embedded in the nuclear envelope which is imaged as flat as possible on the cover glass. The lower and upper ring of all particles are therefore roughly perpendicular to the optical axis of the microscope (Heydarian *et al.*, 2021a).

DNA origami Digits data: The so-called nanoTRON datasets (Auer *et al.*, 2020) consist of DNA origami structures in the shape of the digits 1, 2 and 3 and in the shape of a 3×4 rectangular grid. The data are available online (Heydarian *et al.*, 2021b) and contains on the order of a few thousand particles. These datasets are used to showcase the processing speed advantages of our method.

Simulation data: Simulation data of the DNA-origami TUD-logo was generated as described in Heydarian *et al.* (2018b).

3.2 Parameter settings

A number of parameters in the three algorithmic steps of alignment, classification and connection must be set. The default values given in the table in the Supplementary Appendix C are suitable for most of the cases.

We estimate the number of initial GMM centers K by applying the mean-shift method (Cheng, 1995; Fukunaga and Hostetler, 1975) to the outcome of ζ randomly selected input particles coarsely transformed by JRMPC with K_0 randomly generated GMM centers. We set $K_0 = \min\left((\Sigma_{j=1}^{N} M_j)/N, 100\right)$, i.e. equal to the average number of localizations of all input particles with a minimum of 100. We choose $\zeta = 20$, if the number of input particles $N < 20$ then $\zeta = N$. The value of K estimated in this way is approximately equal to the number of binding sites in most cases. All initial values for the prior probabilities of the K Gaussians are set uniformly to $p_k = 1/K$. The initial values of the center positions μ_k^0 are generated randomly within a rectangular bounding box containing all the localizations. We initialize the transformation as $\mathbf{R}_j^0 = \mathbf{I}_d$ and $\mathbf{t}_j^0 = \overline{\boldsymbol{\mu}}^0 - \overline{\mathbf{v}}_j$, where $\overline{\boldsymbol{\mu}}^0$ is the average of the K GMM centers. The diagonal of the bounding box containing all the input particles after applying the initial translation is set as the initial value of all Gaussian standard deviations σ_k^0. We set the default value for the number of clusters n_l to 2 in the classification step because the registration of JRMPC usually only contains two flipped structures. The threshold ϑ for a cluster to be used in the connection step is set as $N/(n_l + 1)$. The default number of repetitions L for the JRMPC initializations is 2.

We use the default parameter settings throughout with two exceptions. The reconstruction of the nanoTRON 3×4 grid (Fig. 3(d)) uses non-default parameters with a larger number of clusters ($n_l = 8$) to guarantee clusters that contain well-aligned particles. The reconstruction of the 3D NPC particles (Fig. 5) uses a non-default value for the initial Gaussian standard deviation (we use $\sqrt{1000}$, much smaller than the default value) to better fit with the limited range of initial poses of the NPCs. An inferior alignment is observed with the default value. In general, we find that the quality of the individual clusters can be improved by increasing n_l or ϑ. A larger number of JRMPC initializations L can help to increase the number of particles in the final reconstruction after the connection step.

3.3 Benchmark algorithms and evaluation metrics

We compare our proposed method with the all-to-all method (Heydarian *et al.*, 2018b) (Heydarian *et al.*, 2021a). We use the Fourier Ring Correlation (FRC) (Nieuwenhuizen *et al.*, 2013) to measure the resolution of the super-particle reconstructions. We form two independent input image subsets from the super-particle reconstruction to perform the FRC analysis. The first subset is the main cluster $C_m^{l_m}$ and the second subset consists of all other particles in the reconstruction. These two subsets can be used as statistically independent image subsets that are the necessary inputs for the FRC measurement because each subset contains a similar number of different particles from different independent experiments. We cross-checked the outcomes of this FRC computation with the standard method of independently processing two subsets of the total set of input particles and found outcomes within the uncertainty margin of the FRC estimation. In addition, we calculate the localization distribution over the azimuthal angles to analyze the reconstruction symmetry for symmetrical structures. For the 3D NPC data, we also visualize and compare the distribution of z positions of the

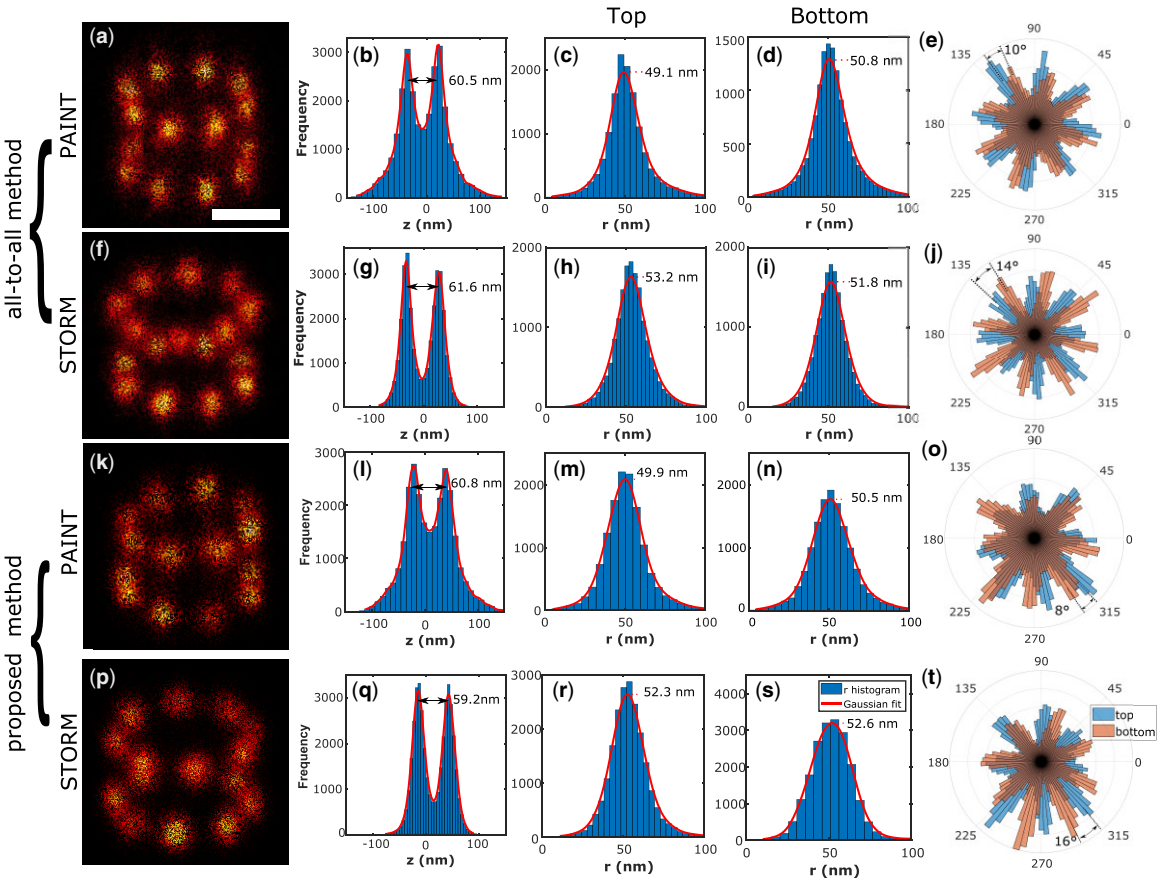

Fig. 5. Comparison of particle fusion performance between our method and all-to-all method on experimental 3D Nup107 particles acquired by different localization microscopy techniques. (**a**) Fusion of 306 Nup107 particles obtained from 3D astigmatic PAINT reconstructed by the 3D all-to-all method. (**b, g, l, q**) Histogram of the z coordinate of localizations in the reconstruction (a). (**c, h, m, r**) Histogram of the radius of upper ring's localizations, (**d, i, n, s**) lower ring. (**e, j, o, t**) Rose plot of the localization distribution over azimuthal angles for the upper and lower rings of the reconstructions. (**f**) Fusion of 356 Nup107 particles obtained from 3D astigmatic STORM reconstructed by the 3D all-to-all method. (**k**) Fusion of 306 Nup107 particles obtained from 3D astigmatic PAINT reconstructed by our method. (**p**) Fusion of 356 Nup107 particles obtained from 3D astigmatic STORM reconstructed by our method. Scale bar indicates 50 nm and applies to a, f, k and p. Rose plots in (e, j) show 8-fold symmetry with nearly equal number of localizations, but symmetry was used here in the reconstruction. Rose plots (o, t) without any prior knowledge reconstruction with our method also shows eight clear peaks however with a stronger variation in the number of localizations

localizations, the radius of each of the two rings, and in a rose plot the localization distribution over azimuthal angles. In the simulations, we compute the root mean square distance between the localizations after particle fusion and the attendant binding sites to quantify the quality of the fusion process (Heydarian *et al.*, 2021a).

4 Results

4.1 Computational cost

Compared to the all-to-all method, which has an unfavorable computational cost scaling as N^2, our method is much faster as it is linear with N. Figure 2(a and b) shows the reconstructions of 383 experimental TUD-logo particles with $DOL = 80\%$ and 788 875 localizations obtained with the all-to-all method and our method. We repeated our method on the 80% DOL TUD-logo particles 30 times in order to assess the uncertainty in FRC-resolution and computation time. Both methods achieve a similar reconstruction quality, consistent with near equal FRC resolutions (3.3 ± 0.3 nm for the single instance of the all-to-all, 3.6 ± 0.3 nm for the 30 runs for our method). The computational time of the all-to-all method, however, is almost 12 times longer than for our method. More importantly, our computational time of 9.6 ± 0.6 min was performed on a simple CPU (40 core Xeon E5-2670v3), opposed to the GPU-implementation of the all-to-all registration (K40c Tesla GPU). The all-to-all method is practically impossible on a CPU when having more than 100 particles. The estimated number of Gaussian centers K is 40 ± 3, which is close to the actual number of binding sites (37). The random initializations of the JRMPC usually result in a final GMM that is similar to the combination of two inverted TUD-logos, which can be classified appropriately in only two clusters.

Our method can effectively handle large amounts of particles because of the favorable reconstruction speed. To show the capability of our method to handle this large data we applied it to the nanoTRON datasets, which contain an order of magnitude more particles than the TUD-logo datasets. We achieved clear structures of the digits 1, 2 and 3 and of the 3×4 grid in only 1.1 h, 1.3 h, 48 min and 4.8 h, respectively, in CPU compared to a computational time of multiple days for the GPU-accelerated all-to-all method. It would have taken several days to resolve the full dataset with the all-to-all method. Due to this speed limitation we only used part of the data in the all-to-all method. The FRC resolution obtained by the all-to-all registration for these four datasets (digits 1, 2 and 3 and of the 3×4 grid) containing 1219, 1309, 1278 and 1194 particles are 3.69 ± 0.02 nm, 4.40 ± 0.19 nm, 3.98 ± 0.22 nm and 3.59 ± 0.15 nm, respectively (Huijben *et al.*, 2021). Our reconstructions include 4155, 4943, 2541 and 5961 particles for these four datasets and the FRC resolutions are 2.76 ± 0.92 nm, 2.80 ± 0.54 nm, 3.21 ± 0.33 nm and 3.51 ± 0.28 nm, respectively. These numbers are smaller as we are able to assemble more particles in the final reconstruction compared to the all-to-all method. For the digits 1, 2, and 3, the estimated K (25, 23, 34) is close to the actual number of binding sites (18, 23, 25). For the 3×4 grid particles, our K-estimation algorithm estimates $K = 42$ which is much more

than the 12 binding sites. For that reason the JRMPC reconstructions have more clusters and we need a larger $n_l = 8$ to separate them correctly.

4.2 2D NPC data: influence of symmetry

Our method also overcomes the second disadvantage of the all-to-all method, the hot-spot problem occurring for symmetrical structures. In Figure 4, we compare reconstructions of 2D NPC particles with 8-fold rotational symmetry. The reconstruction of the all-to-all method without prior knowledge (Fig. 4(a and b)) shows one apparent 'hot-spot' with more than 600 localizations compared to other blobs with around 400 localizations. After imposing eight-fold rotational symmetry the hot-spot disappears (Fig. 4(c)). Imposing this symmetry changes the ellipticity of the reconstructed NPC ring from the earlier 0.89 to 0.99. So, symmetry has been restored, but at the expense of a shape that changed from an ellipse to a circle. Our method applied to the same NPC particles does not result in a hot-spot (Fig. 4(e)), quantified by a more uniform distribution of localizations over the 8 peaks (compare (b) and (f)). The ellipticity of our reconstruction is 0.86 which matches reasonably well with the all-to-all value of 0.89. The number of Gaussian components K in the GMM is estimated by our algorithm to be 8 which is obviously equal to the number of visible binding sites in the 2D NPC.

4.3 Low labeling 2D DNA origami data

A major accomplishment of the all-to-all method is its ability to handle poorly labeled data. It appears our method outperforms the all-to-all method even in this respect. Figure 2(c–f) shows a comparison of reconstructions of hundreds of TUD-logos with low DOL values equal to 50% and 30%. Our method results in a visually better reconstruction quality, especially for the worst quality $DOL = 30\%$ dataset (compare Figure 2(e and f)). Nearly all binding sites on the origami at a distance of about 5 nm are resolved in (f) where in (e) especially the edges are washed out and localizations are concentrated to a few binding sites. This is consistent with the FRC resolutions of 3.1 nm and 3.3 nm for the 50% and 30% DOL datasets, respectively, which compares favorably with the FRC resolutions for the all-to-all method equal to 3.5 nm and 5.0 nm for the 50% and 30% DOL datasets, respectively. The mean-shift method estimates $K = 46$ for the data with 30% DOL and $K = 37$ for 50% DOL. These two K values are very close to the actual number of 37 binding sites of the origami design. The initial Gaussian standard deviation is quite large (~100 nm) at first. Most of the Gaussian components shrink to a small size (less than 3 nm) eventually, and only a few to a medium size (~10 nm). Most of the initially randomly generated GMM centers $\boldsymbol{\mu_k}$ are finally positioned near the binding sites of the TUD-logo.

4.4 3D NPC data

Another major achievement of the all-to-all method is the ability to reconstruct 3D data (Heydarian *et al.*, 2021a). Our method shows a comparably good performance on 3D datasets. Figure 5 shows a comparison of 3D Nup107 NPC structures imaged with both PAINT and STORM. Our method shows reconstructions of similar quality as the all-to-all method (compare Figure 5(a and k) and compare Figure 5(f and p)). Here, the all-to-all method relies on detecting the rotational symmetry from the data and subsequently promoting the symmetry in the reconstruction. In contrast, neither prior knowledge or detection of symmetry nor extra post-processing is needed with our method. Comparison of Figure 5(b, g, l, q), (c, h, m, l) to (d, i, n, s), respectively, further shows that our method obtains similar NPC structural parameters (the distance between the nuclear and cytoplasmic rings and their radius) as the all-to-all method. The rose plots Figure 5(e, j) obtained from the all-to-all method's reconstructions show 8-fold symmetry for each ring, and the number of localizations in each peak is almost the same. The rose plots Figure 5(o, t) of our reconstructions also clearly show eight peaks for each ring, but the number of localizations in each peak is slightly different. This is reasonable considering that our method does not rely on symmetry in the reconstruction. Our K-estimation algorithm estimates $K = 34$ for both cases, which is also reasonable as the number of actual binding sites should be 32 given the structure of the EM model (Kosinski *et al.*, 2016; Thevathasan *et al.*, 2019). The default value of σ_k does not work here and we used $\sigma_k^0 = \sqrt{1000}$ nm instead. The final center points of the GMMs are nearly all distributed inside the 16 spheres of the 3D NUP reconstructions.

4.5 Simulation data

We explore the limitations of the proposed method in terms of DOL, localization precision and the number of particles by applying our method on simulated TUD-logo datasets. Further details and results can be found in Supplementary Appendix D. Even with poor quality datasets (DOL as low as 40%, or localization precision as low as 12 nm, or a number of particles as low as 10), our method can obtain reconstructions with registration error in the range 5–10 nm.

5 Discussion

Several of the results we obtained can be qualitatively understood: in comparison to the all-to-all-method our approach produces better results for poor, underlabeled data. The reason is that in the pairwise registration of the all-to-all method pairs of poor quality particles must be aligned, which is more error prone than our approach where each of the poor quality particles is aligned to the average of all particles. The same line of reasoning applies to the case of symmetric structures. The pairwise registration of the all-to-all method aligns random peaks that occur through the stochastic variations of labeling within the particles, while for our approach each particle is aligned to the average of all particles which smoothens out the stochastic variations in labeling.

We attribute the JRMPC local optima that consist of several distinct clusters with different poses to a difference in convergence rate between the widths of the Gaussian components and the particle rotations. It seems that the Gaussian widths shrink relatively fast, while the particle rotations only change slowly, as the iteration progresses. This results in posterior probabilities α_{kij} for the Gaussian component k that is nearest to localization i of particle j that quickly converge to nearly one and to virtually zero for the other Gaussian components. On the other hand, for the case of 3D NPC particles with a limited range of poses in the dataset, the widths of the Gaussian components appear too large, leading to sets of particle rotations that are distributed too broadly. Summarizing, the reconstruction quality appears to be sensitive to the initial setting and convergence rate of the Gaussian widths.

Next to the limitations of our method on DOL and localization uncertainty assessed by the simulation study, there are also some assumptions that go into the proposed method that we wish to emphasize now. Firstly, the underlying biological structure is assumed to be sufficiently rigid for the overall averaging to make sense. Secondly, we assume a single underlying structure. Recently, however, we have studied the detection of structural heterogeneity using particle fusion methods (Huijben *et al.*, 2021). Our method can also vastly accelerate the workflow of Huijben *et al.* (2021), as there the initial step is to find a global alignment of all particles in the dataset, followed by quantification of pairwise (dis)similarity between the particles. Thirdly, the idea of fitting a GMM to localization microscopy data sets matches well structures with a discrete number of binding sites for fluorophores and corresponding datasets with multiple localizations per binding site. In that case the GMM centers will tend to gravitate toward the different binding sites. In case there are only a few localizations per binding site, however, it will be difficult to match the GMM centers to the binding sites, and the quality of the registration process may be compromised.

A number of algorithmic improvements can be envisioned. First of all we could incorporate the localization uncertainties in the JRMPC method, such that the probability of localization i of particle j to fit Gaussian component k is a normal distribution with a variance that is the sum of the variance due to the localization uncertainty and the variance of the Gaussian component. Especially in cases where the localization uncertainty is on the order of the

distance between binding sites, or where there is a broad distribution of localization uncertainties, or when the localization uncertainty is anisotropic (for 3D datasets), this may improve the sensitivity to the initial setting of the widths of the Gaussian components, as well as promote convergence to a global optimum. Another improvement may be found in a better description of the quality of the clusters. Now we opt for the simple criterion of number of particles in the cluster. Using the FRC resolution may be a better practice for assessing cluster quality.

6 Conclusion

We have proposed a fast particle fusion method with computational complexity that scales linearly with the number of input particles. In our method, we apply the JRMPC method for multiple initializations and then use classification and connection steps to generate a correct reconstruction with as many particles as possible. The reconstruction quality of our method is measured by the FRC resolution and compared with the all-to-all method, revealing that our results are of comparable or better quality. Our method is fast, even without GPU acceleration, avoids symmetry artifacts, applies to 2D and 3D datasets, and reconstructs poor data with a limited number of particles, a low density of labeling and a large localization uncertainty.

Acknowledgements

We thank Sabri Bolkar for initial attempts to apply the JRMPC method to single-molecule localization microscopy.

Funding

This work has been supported by the Dutch Research Council (NWO), VICI grant no. 17046 for B.R. and W.W.

Conflict of Interest: none declared.

References

Auer,A. *et al.* (2020) nanoTRON: a picasso module for MLP-based classification of super-resolution data. *Bioinformatics*, **36**, 3620–3622.

Broeken,J. *et al.* (2015) Resolution improvement by 3D particle averaging in localization microscopy. *Methods Appl. Fluoresc.*, **3**, 014003.

Cheng,Y. (1995) Mean shift, mode seeking, and clustering. *IEEE Trans. Pattern Anal. Mach. Intell.*, **17**, 790–799.

Evangelidis,G.D. and Horaud,R. (2017) Joint alignment of multiple point sets with batch and incremental expectation-maximization. *IEEE Trans. Pattern Anal. Mach. Intell.*, **40**, 1397–1410.

Fox-Roberts,P. *et al.* (2017) Local dimensionality determines imaging speed in localization microscopy. *Nat. Commun.*, **8**, 1–10.

Fukunaga,K. and Hostetler,L. (1975) The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Trans. Inform. Theory*, **21**, 32–40.

Gray,R.D. *et al.* (2016) Virusmapper: open-source nanoscale mapping of viral architecture through super-resolution microscopy. *Sci. Rep.*, **6**, 29132.

Hell,S.W. (2009) Microscopy and its focal switch. *Nat. Methods.*, **6**, 24–32.

Heydarian,H. *et al.* (2018a) Single-molecule localization microscopy (SMLM) 2D TU Delft logos. *4TU.ResearchData. Dataset*. doi:10.4121/uuid:0d42a28f-f625-41a3-ba77-25e397685466.

Heydarian,H. *et al.* (2018b) Template-free 2D particle fusion in localization microscopy. *Nat. Methods*, **15**, 781–784.

Heydarian,H. *et al.* (2021a) 3D particle averaging and detection of macromolecular symmetry in localization microscopy. *Nat. Commun.*, **12**, 1–9.

Heydarian,H. *et al.* (2021b) Single-molecule localization microscopy (SMLM) 2D digits 123 and TOL letters datasets. *4TU.ResearchData. Dataset*. doi:10.4121/14074091.v1.

Heydarian,H. *et al.* (2021c) Single-molecule localization microscopy (SMLM) 3D datasets. *4TU.ResearchData. Dataset*. doi:10.4121/13797686.v1.

Huijben,T.A.P.M. *et al.* (2021) Detecting structural heterogeneity in single-molecule localization microscopy data. *Nat. Commun.*, **12**, 1–8.

Jain,A.K. *et al.* (1999) Data clustering: a review. *ACM Comput. Surv.*, **31**, 264–323.

Klein,T. *et al.* (2014) Eight years of single-molecule localization microscopy. *Histochem. Cell Biol.*, **141**, 561–575.

Kosinski,J. *et al.* (2016) Molecular architecture of the inner ring scaffold of the human nuclear pore complex. *Science*, **352**, 363–365.

Löschberger,A. *et al.* (2012) Super-resolution imaging visualizes the eightfold symmetry of gp210 proteins around the nuclear pore complex and resolves the Central channel with nanometer resolution. *J. Cell Sci.*, **125**, 570–575.

Mead,A. (1992) Review of the development of multidimensional scaling methods. *J. R. Stat. Soc. Series D Stat.*, **41**, 27–39.

Metzger,M. *et al.* (2016) Resolution enhancement for low-temperature scanning microscopy by cryo-immersion. *Opt. Express*, **24**, 13023–13032.

Nieuwenhuizen,R.P. *et al.* (2013) Measuring image resolution in optical nanoscopy. *Nat. Methods*, **10**, 557–562.

Rieger,B. and Stallinga,S. (2014) The lateral and axial localization uncertainty in super-resolution light microscopy. *Chemphyschem*, **15**, 664–670.

Szymborska,A. *et al.* (2013) Nuclear pore scaffold structure analyzed by super-resolution microscopy and particle averaging. *Science*, **341**, 655–658.

Thevathasan,J.V. *et al.* (2019) Nuclear pores as versatile reference standards for quantitative superresolution microscopy. *Nat. Methods*, **16**, 1045–1053.

Vicidomini,G. *et al.* (2018) STED super-resolved microscopy. *Nat. Methods*, **15**, 173–182.