A marker-free automatic alignment method based on scale-invariant features

Renmin Han\textsuperscript{a,b}, Fa Zhang\textsuperscript{a,e}, Xiaohua Wan\textsuperscript{a}, Jose-Jesus Fernández\textsuperscript{c}, Fei Sun\textsuperscript{d}, Zhiyong Liu\textsuperscript{e}

\textsuperscript{a}Key Lab of Intelligent Information Processing and Advanced Computing Research Lab, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China
\textsuperscript{b}University of Chinese Academy of Sciences, Beijing, China
\textsuperscript{c}National Centre for Biotechnology, National Research Council (CNB-CSIC), Campus UAM, C/Darwin 3, Cantoblanco, 28049 Madrid, Spain
\textsuperscript{d}Institute of Biophysics, Chinese Academy of Sciences, Beijing 100101, China
\textsuperscript{e}State Key Lab for Computer Architecture, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China

\section*{A R T I C L E   I N F O}

Article info:
Received 15 September 2013
Received in revised form 17 February 2014
Accepted 18 February 2014
Available online 27 February 2014

Keywords:
SIFT
Electron tomography
Alignment
Bundle adjustment

\section*{A B S T R A C T}

In electron tomography, alignment accuracy is critical for high-resolution reconstruction. However, the automatic alignment of a tilt series without fiducial markers remains a challenge. Here, we propose a new alignment method based on Scale-Invariant Feature Transform (SIFT) for marker-free alignment. The method covers the detection and localization of interest points (features), feature matching, feature tracking and optimization of projection parameters. The proposed method implements a highly reliable matching strategy and tracking model to detect a huge number of feature tracks. Furthermore, an incremental bundle adjustment method is devised to tolerate noise data and ensure the accurate estimation of projection parameters. Our method was evaluated with a number of experimental data, and the results exhibit an improved alignment accuracy comparable with current fiducial marker alignment and subsequent higher resolution of tomography.

\textcopyright{} 2014 Published by Elsevier Inc.

\section*{1. Introduction}

Electron tomography (ET) is a promising technology that allows the three-dimensional imaging of cellular ultrastructure. The structure is reconstructed from a tilt series of micrographs taken at different orientations. However, transformation and deformation of the sample are inevitable when the sample is tilted along a fixed axis. To obtain high-quality reconstructed results, accurate alignment is critical before reconstruction.

There are two types of alignment methods, fiducial marker-based alignment and marker-free alignment. Fiducial marker-based alignment is currently the most accurate alignment method. Unfortunately, fiducial markers are not always accessible, because sometimes it is impossible to have gold beads embedded in a sample and sometimes it is difficult to find enough gold beads at the region of interest. Moreover, the use of colloidal gold may interfere with the sample and introduce undesirable artifacts. Additionally, the selection of markers is usually manual and very time-consuming. In contrast, marker-free alignment does not require fiducial markers. It can be subdivided into two categories of methods, correlation methods and feature-based methods. Correlation methods, such as cross-correlation (Guckenberger, 1982) and common lines (Liu et al., 1995), have been widely used in coarse alignments to solve large translation or in-plane rotation problems. However, these methods neglect motion in real space and result in accumulated correlation errors. To compensate for these shortcomings, Winkler and Taylor (2006) proposed a solution combining cross-correlation with a reconstruction reprojection method, but this method consumes excessive computational resources. Compared with correlation methods, feature-based methods provide a model that is closer to the real conditions and is not computationally intensive. Feature-based methods utilize image features as virtual markers and align images by minimizing the reprojection error of virtual markers. The principle of feature-based alignment is same as that of marker alignment. Usually, features can be determined by informative surroundings, such as a Harris Corner (Brandt et al., 2001; Brandt and Ziese, 2006), Canny edge, or contour line (Phan et al., 2009), and appointed landmarks (Sorzano et al., 2009). Castaño-Díez et al. (2007, 2010) used image patches instead of corner points in the proceeding cryo-ET series. These features are iteratively tracked from the corresponding area of adjacent images according to the normalized cross-correlation. However, such a tracking method is not always robust, especially in the case of low Signal-to-Noise Ratio (SNR) where there are insufficient distinguishable gray levels. These methods mainly
introduce two types of errors that are highly influential with respect to geometry parameter determination, feature localization error and false matching. Although feature-based methods are more accurate than correlation methods, three key issues remain to be resolved. The first is how to accurately detect and locate interest points (features). As mentioned above, virtual markers are extracted with a computer vision technique, which inevitably introduces localization errors. The second is how to partition a huge amount of features into tracks. Typically, an alignment operation involves hundreds of images. For every image, thousands of features could be extracted. Tracking hundreds of thousands of features is very time-consuming. In addition, marker mismatching and matching collisions must be resolved during tracking. The third issue is how to optimize projection parameters after tracking. It is very difficult to solve this problem because there are a large number of tracks and the length of each track is relatively short compared to the size of the image stack. Because not all of the tracks are consistent, a robust method for parameter optimization must be used.

To overcome these problems, we propose a new marker-free method based on Scale-Invariant Feature Transform (SIFT) to solve marker-free alignment. SIFT is a well-known technique used in computer vision that can locate points in scale-space and utilize redundant feature information. Compared to other previous methods, our method has several advantages. First, we utilize SIFT to detect and recognize a huge number of significant virtual markers that are invariant to the changes of scale, orientation, noise etc. After the “significant” or “interesting” points are detected, we focus only on the important parts of the tilt series, which potentially ignores background areas. Second, in addition to the detection of localization, the distinctive information of features is characterized, which makes feature matching and tracking more robust. Furthermore, our method contains an effective tracking model to make feature tracking more efficient and resistant to dubious matching. Third, in contrast to previous methods using a simplified version of the affine model (for example, the research in Brandt et al., 2001; Brandt and Ziese, 2006), our method uses a more parameterized model, which benefits from its high-quality tracking and can make further analysis of tilt series possible, resulting in more accurate alignment. Experimental datasets were tested and proved that our alignment method can optimize the parameters with subpixel accuracy of the reprojection residual.

The remainder of the paper is organized as follows. In Section 2, we introduce the framework of our method. First, we introduce the usage of SIFT in electron micrographs and demonstrate our effective matching strategy and tracking model. Then, an incremental bundle adjustment procedure designed especially for our approach is proposed. In Section 3, we present our experimental results and analysis. Section 4 is focused on discussion and conclusion.

2. Method

Our method consists of four steps. The first step is to extract the precise location and descriptions of features from projection images. We utilize SIFT to obtain subpixel feature localizations and descriptor vectors with redundant information, which ensures that the extracted features are invariant to scale, rotation and illumination changes. The second step is to match corresponding features. Because of the large number of feature points, we propose a location-based search method to ensure accuracy and accelerate the matching speed. The third step is to track matching pairs consistently across the tilt series. Based on the transitivity of matched peers, here we first develop a matching strategy to reduce the matching cost and then propose a novel tracking model to reduce tracking complexity. The final step is to optimize the projection parameters with the configured tracks and, if necessary, to geometrically transform the images. We first present the parameter optimization model of our method, and then propose an incremental bundle adjustment method to solve the optimization problem. Our approach obtains results with improved accuracy which is comparable with that got by fiducial marker alignment.

2.1. Feature extraction with SIFT

One feature is composed of two parts, the location and distinctive information (descriptor). In previous feature-based methods, only the gray values in the neighborhood are considered for cross-correlation, and the abundant information that the neighborhood renders is neglected. Thus, such processes are apt to mismatch and do not generate high-quality tracks. Our method utilizes the SIFT detector (Lowe, 2004) to extract features. SIFT can localize the most stable points in images and form the neighborhood information into a 128-dimensional descriptor that contains gradient and magnification information in a redundant manner. In fact, SIFT has been widely used in low SNR image analysis, for example, Alzheimer’s disease detection in medicine (Toews et al., 2010) and image stitching in ET (Kaynig et al., 2010). These reports showed a high accuracy of localization and discrimination of detail of SIFT. Mikolajczyk and Schmid (2005) compared the SIFT descriptor with other invariant feature descriptors and drew the conclusion that SIFT performed the best under the changes of scale, rotation and illumination.

SIFT consists of the following four major stages: (1) Scale-space extrema detection. (2) Keypoint localization. (3) Orientation assignment. (4) Keypoint descriptor.

Scale-space extrema detection is to identify the locations which appear repeatedly for the same object in different views and scales. To detect locations that are invariant to the scale change of images, we search for stable features across all possible scales using a continuous scale function known as scale space. The scale space is a collection of the image function convolved with various Gaussian kernels, which is defined as the function $L(x,y;\sigma)$:

$$L(x,y;\sigma) = \frac{1}{2\pi\sigma^2} e^{-\frac{(\sqrt{\Delta}x^2+\Delta y^2)}{2\sigma^2}} + I(x,y)$$  \hspace{1cm} (1)$$

where $I(x,y)$ is the input image, $*$ is the convolution operation in $x$ and $y$, $\sigma$ is the Gaussian scale and various values of $\sigma$ will produce different scale in the scale space. To efficiently detect stable key locations in scale space, Lowe (1999) adopted scale-space extrema in the difference-of-Gaussian (DoG) function, defined as $D(x,y;\sigma)$, which can be computed from the difference of two nearby scales separated by a constant multiplicative factor $k$:

$$D(x,y;\sigma) = (G(x,y;\sigma) - G(x,y;\sigma)) * I(x,y) = L(x,y;\sigma) - L(x,y;\sigma)$$  \hspace{1cm} (2)$$

The DoGs of various Gaussians compose the DoG space. DoG is an approximation of the scale-normalized Laplacian-of-Gaussian, which is required for true scale invariance (Lindeberg, 1994). An extremum point (maxima or minima of the DoG images which is detected by comparing a central pixel to its 26 neighbors in 3 × 3 regions at the current and adjacent scales) of scale-space can produce the most stable features for one image (Mikolajczyk, 2002).

Keypoint localization is to fit a candidate with its adjacent data according to the location, scale, and ratio of the principal curvatures. Candidate points that have low contrast or are poorly localized along the edges (points with peak value in the DoG but sensitive to noise) will be rejected in this stage. This process provides a substantial improvement in the stability of features. Orientation assignment is to assign one or several orientations to each keypoint based on gradient directions. Orientations are
determined by calculating a histogram of gradient orientations of
the keypoint’s neighborhood. By assigning a consistent orientation
to each keypoint, we can construct a canonical view of a certain
feature that is invariant to image rotation.

The last stage is to generate the descriptor by sampling the
magnitudes and orientations of the image gradient of the area
around the keypoint in the scale space. In practice, the area will
be adjusted according to the major orientation of keypoint and is
divided into several subareas. Each subarea overlaps with the
neighbors. Such redundant information guarantees that the
descriptor is invariant to the change of scale, but only approxi-
mately invariant to small change of affine transformation. Finally,
a 128-dimension vector is formed as a descriptor, which is then
normalized to a magnitude of unity so that there is no variance
due to illumination.

Fig. 1 illustrated a small mitochondrial image patch with key-
points in it. In this figure, vectors originate from keypoint locations
representing the dominant direction of the SIFT descriptor as well
as their strength for each keypoint. As shown in Fig. 1, keypoints in
this image are located around membranes and other significant
changes. Typically, approximately 4000 features can be ex-
tracted from a 1024 × 1024 image, which is sufficient for the deter-
mination of geometry.

2.2. Feature matching

Algorithm 1. Match features of two images

Initialization: Set $S = \{ p_{ij} | j = 0, 1 \ldots m \}, S' = \{ p_{ij}' | j = 0, 1 \ldots n \}, M = \emptyset$
1: for all point $p_{ij} \in S$ do
2: get subset $W = \text{Window}(p_{ij}, S')$
3: for all point $p_{ijk} \in W$ do
4: if scale or orientation of $p_{ijk}$ is in range of $p_{ij}$’s then
5: $d(p_{ij}, p_{ijk}) = \text{Euclidean distance}(p_{ij}, p_{ijk})$
6: if $d(p_{ij}, p_{ijk})$ is smaller than the minimum or sub-
minimum then
7: update the minimum or sub-minimum distance
8: end if
9: end if
10: end for
11: get the minimum distance $d_{\text{min}}$ and sub-min $d_{\text{sub}}$
12: if $d_{\text{min}} < \text{threshold} - d_{\text{sub}}$ then
13: $M = M \cup \{ p_{ij}, W_{\text{minde}} \}$
14: end if
15: end for
16: RANSAC of $M$
17: return $M$

Three aspects should be considered in feature matching. The
first aspect is the robustness of feature matching. As mentioned
in the previous section, although classical normalized cross-
correlation is a good measure of similarity, the descriptor vector of
SIFT is more stable under the change of image rotation and affine
transformation (Mikolajczyk and Schmid, 2005). Furthermore, to
match features between two images, a global search is not feasible
in our case, because most biological samples are symmetric or have
some parts similar to each other, which leads to unpredictable mis-
matching. The second aspect is the pruning of mismatching and
dubious matching. Mismatching and dubious matching are inevita-
able due to low SNR, variation of the sample during tilt and biolog-
cal self-similarity. The third aspect is the matching complexity. A
typical global search method for multi-dimensional vectors is
based on a kd-tree (Moore, 1991). The complexity of building a
kd-tree differs according to its implementation, which generally
refers to $O(kn \log n)$. The search complexity is $O(m \log n)$ on average.
However, the $k$ here is 128, referring to the dimension of the
descriptor, and $n$ is about 4000–5000, which makes feature match-
ing very time-consuming.

To solve the above problems, here we propose a location-based
search technique, described in Algorithm 1, where $S$ is the set of
keypoints of the $i$th image, and $S'$ is the set of keypoints of the
$j$th image. $\text{Window}(p_{ij}, S')$ is a window function that returns the
points in set $S'$ and near $p_{ij}$. $d(p_{ij}, p_{ijk})$ is the Euclidean distance
of the descriptors, which is used to distinguish the similarity be-
tween two features, and $\text{threshold}$ is the ratio which indicates their
similarity.

The feature descriptor intrinsically contains scale and redu-
dant information of the neighborhood and is invariant to change
of rotation. And the Euclidean distance comparison of the descrip-
tor vector can provide high accuracy for distinction (Mikolajczyk
and Schmid, 2005). However, the selection of the threshold ratio
value must be determined experientially. The larger the ratio is,
the more matching pairs are obtained, but with less confidence.
Based on our experience, we suggest a ratio between 0.4 and 0.6.
This procedure implicitly eliminates spurious matching.

High-quality tracking is critical to be able to cope with mis-
matching and dubious matching. To prune mismatching, we utilize
epipolar constraint. The epipolar constraint provides a point-line
relationship between two geometry transformations represented
by a $3 \times 3$ matrix $F$ as $m'Fm = 0$ for every matching pair $m$ and
$m'$ (in homogeneous form). To obtain unknown constraints, RAN-
SAC (Random Sample consensus) (Fischler and Bolles, 1981), a very
robust and efficient method of removing outliers, is used. The main
process is as follows:

- initially, choose 8 point pairs and calculate their epipolar
  geometry;
- set a threshold and check each match with the obtained funda-
  mental matrix. Matching pairs whose error is within the thresh-
  old are accepted to form a consensus set;
- repeat to get several $F$ matrices and the corresponding consen-
  sus sets until terminal conditions are reached;
- find the maximum consensus set, then re-output a fundamental
  matrix and the corresponding matching pairs.

The inlier matching pairs are directly used in tracking, and the $F$
matrix provides a criterion to determine dubious matching that is
used in the next section.

In our location-based search method, the location-based win-
dow function is used to reduce the search scope of the matching
operation. Our window function is supported by a 2D data struc-
ture, which guarantees a fast insert and search operation in
$O(\log n)$. A window width of 5% of the maximum image width is

![Image](https://example.com/image.png)
appropriate. The calculation cost between $p_{ij}$ and subset $W$ is significantly smaller than that of the kd-tree considering that $k$ is 128 and $n$ is large.

2.3. Feature tracking

The classical way to extend the tracks is to search for and determine the corresponding matched features in every view, then assign them to the appropriate track. It means that for each image, we need to search all of the other images to determine the matched features. Such a solution is only suitable for a small number of feature pairs and views. However, in our case, the number of images in a tilt series and the number of extracted features in each image are both very large. Feature tracking in the classical way would be very time-consuming.

However, our analysis indicated that for a given image, we do not need to match all the other images in a tilt series. Here, we used the number of long tracks (obtained through over 15 images) as a criterion and statistically analyzed what percentage of long tracks could be covered if a subset of adjacent views in the tilt series were matched. If the range is $x$, the $i$th image will be matched with the $(i + 1)$th, $(i + 2)$th, ..., and $(i + x)$th images. The procedure in which we match every two images in the tilt series and combine matching pairs into tracks is called full-matching, and the procedure in which we only match a range of adjacent images and add the matching pairs into tracks is called adjacent-matching. The experimental results are illustrated in Fig. 2, where the $x$-axis indicates the range of adjacent views, and the $y$-axis indicates the percentage of the number of long tracks obtained by adjacent-matching compared to full-matching. From Fig. 2, we found that over 80% of the long tracks are identified if the range of adjacent views is three, and almost 100% are identified if the range is 10.

Based on the above analysis, two techniques are developed to solve the tracking problem: a matching strategy to reduce the matching cost and a tracking model to reduce the tracking complexity.

Matching strategy: In the beginning, we should ensure two premises. First, the SIFT descriptor is stable and reliable, so a matched feature appearing in one image would have a corresponding peer in another. Secondly, there is small change in a series of gradual tilts. These characteristics have been confirmed by the above discussion. For a distinct feature in one image, we can normally find the corresponding feature in its neighbor images. For example, for features in the $n$th image, its peers mostly appear in the $(n - 1)$th, $(n + 1)$th, $(n + 2)$th image and so on. If we want to find the matching pairs between the $n$th and $(n + 2)$th image, we need to match features between the $n$th and $(n + 1)$th, the $(n + 1)$th and $(n + 2)$th images, and then combine the results. This transitivity of the image in a tilt series has been widely used in electron tomography (Brandt and Ziese, 2006).

Due to noise or deformation in electron tomography, transitivity may not be ideal, as we discussed above. Thus, we propose the following matching strategy:

1. Initialize $step = 1$, while $step \leq MAXSTEP$;
2. For the $n$th image of series, match features between the $n$th and $(n + step)$th feature sets ($n = 1, 2, ..., n$);
3. $step = step + 1$;
4. Repeat until condition in 1) is not satisfied.

For the sake of the completeness of information, we should generate tracks that are as long as possible, i.e., MAXSTEP as large as possible. However, this approach is contrary to our purpose. Through the analysis of SIFT’s behavior in a large number of samples, we conclude that SIFT has enough stability to ensure the transitivity of feature peers, as shown in Fig. 2. We found that a MAXSTEP value of 3–8 is enough to ensure the feature’s transitivity.

Tracking model: It is important to improve the extending of tracks, which is a topic that has not been discussed in previous reports. The main reason is that the number of points involved in previous studies is relatively small. In our case, two aspects of tracking need to be solved. The first is how to search points quickly. Binary searching in an ordered sequence is a good option, but it ignores the inherent local constraints in a matching set. The second aspect is how to cope with dubious matching (or matching collision). This issue is important to ensure the consistency of a track.

We devised a tracking model to solve the problems described above. The main idea is arranging a feature point into the place where it should be, instead of adding a feature to tracks explicitly. We use a 2-dimensional data structure, called point-plane, to store the feature points of an image. The structure could be implemented in a balanced binary tree, which guarantees successful searching and inserting a point in $O(\log n)$ complexity. The matched pairs in different point-planes are linked to form a 3-dimensional space (called tracking space). Once the tracking space has been constructed, the cost of visiting tracks is constant. In order to resolve the problem of matching conflict, the error of the epipolar constraint, i.e., $MTm$, is utilized. The dubious pair with the larger epipolar error of the conflicted pairs will be pruned. Fig. 3 shows an example of our solution.

2.4. Parameter optimization

To solve the projection parameters, we need a suitable projection model. For transmission electron microscopy (TEM), an affine model is adaptable. The projection model is formulated as follows:

\[
\begin{bmatrix}
    u \\
    v \\
    z_c
\end{bmatrix} = K(I, 0) \begin{bmatrix}
    R & t \\
    0 & 1
\end{bmatrix} \begin{bmatrix}
    X \\
    Y \\
    Z
\end{bmatrix}
\]

where $K$ is the intrinsic matrix of the camera, $(I, 0)$ is the projection matrix, $R$ is an $3 \times 3$ orthogonal matrix, which can also be written as $R_x \cdot R_y \cdot R_z$, representing rotation in space,

\[
R_x = \begin{bmatrix}
    1 & 0 & 0 \\
    0 & \cos \alpha & \sin \alpha \\
    0 & -\sin \alpha & \cos \alpha
\end{bmatrix}
\]

\[
R_y = \begin{bmatrix}
    \cos \beta & 0 & -\sin \beta \\
    0 & 1 & 0 \\
    \sin \beta & 0 & \cos \beta
\end{bmatrix}
\]
When generally, the intrinsic matrix rotation matrix electron tomography changes minimally during the experiment. The parameter optimization is a non-linear problem. A significant difference between this parameter optimization and other optimization problems is that there is a large number of projecting points involved, while relatively few parameters of projection matrices and spatial points need to be estimated, which makes general optimization methods, such as gradient decent and regression, unsuitable in these conditions.

In our solution, we consider bundle adjustment (Triggs et al., 2000) to solve this problem. Bundle adjustment has been widely used in computer vision scope, but it is not feasible in our case. Although epipolar constraint is used to remove mis-matches, there are still inconsistent points (outliers) in the tracks, and bundle adjustment is sensible to outliers. Furthermore, although we got thousands of tracks by feature tracking, most of the tracks are not long enough to cover all images, and bundle adjustment requires that tracks go through as many views as possible. Thus, we propose an incremental bundle adjustment for parameter optimization.

In the projection parameter optimization, three problems need to be considered. First, how to optimize a problem with a large amount of input points and relatively few projection parameters. Second, how to reduce the adverse impact of outliers in every iteration of bundle adjustment. Third, how to implement bundle adjustment in our condition where the length of tracks is relatively short. To overcome these problems, three key techniques are proposed in our incremental bundle adjustment. Our method is summarized in Fig. 4.

(a) **Solving parameters:** The optimization problem can be redefined as follows. Let all parameters (projection parameters and spatial points) be converted into vector \( p \in \mathbb{R}^m \), and let \( f \) be a mapping that maps \( p \) to a vector \( x \in \mathbb{R}^n \), \( x = f(p) \). Thus, \( f \) denotes the projection operation, and \( x \) denotes all the predicted projection points (vector \( x \in \mathbb{R}^n \) denotes the measured projection points).

The cost function \( E \) can be formulated as \( E = \sum_j || e_j ||^2 \), where \( e = || x - \hat{x} || \) (\( \| \cdot \| \) denotes L2-norm) is the residual. Obviously, the original problem has been formulated in a nonlinear least-square style. The Levenberg–Marquardt (L-M) algorithm (Kanzow et al., 2004) is utilized to solve the parameter optimization problem. The L-M algorithm approximates \( f \) in its neighborhood \( p \) as \( f(p + \sigma_p) \approx f(p) + J f \sigma_p \). When \( || \sigma_p \|| \) is small, \( J \) is the Jacobian matrix of \( f \). Note that the number of measured projection points is much larger than that of calibrated parameters. Thus, \( m \ll n \), and the Jacobian matrix \( J \) is sparse. Taking advantage of this sparse property can reduce computational complexity enormously, and in our solution, a general sparse bundle adjustment package (Lourakis and Argyros, 2009) is used for convenience. Details are discussed in the Appendix.

(b) **Adjusting outliers:** For the sake of robustness, after every iteration of optimizing the parameters, we reproject the spatial points \( \hat{x}_{ij} \) to \( 0, 1, \ldots \) and calculate reprojection residual
3. Experiments and results

3.1. Test dataset

Three samples of different types are presented to demonstrate the performance of the proposed method. Two of the samples were collected without fiducial markers, provided by the Institute of Biophysics, Chinese Academy of Sciences. For the sake of comparison, one is thin and has a small tilt-axis shift, which allows acceptable performance of a correlation-based alignment. The other has a relatively high thickness and tilt-axis shift, for which a correlation-based method is unsuitable. The third sample was downloaded from the IMOD homepage and was embedded with well-distributed fiducial markers. For our convenience, the cross-correlation alignment provided by IMOD (Kremer et al., 1996) was carried out in advance.

The first sample is a tilt series of mitochondria of mouse hepatic cells without fiducial markers. Projection images from −52.0° to +59.0° at 1° intervals were taken with a FEI Tecnai 20, with voltage at 200 kV. The images are 2048 × 2048 in pixel size and 0.4 nm in pixel width. A piece of the images is shown in Fig. 5(a). The sample is thin and has minimal tilt-axis shift, which allows a convenient comparison of the results of our method and the cross-correlation alignment provided by IMOD.

The second sample is a tilt series of the caveola of porcine aorta endothelial cells (PAE cells), which is also lacking fiducial markers. Projection images from −60.0° to +58.0° at 1° intervals were taken with a FEI Tecnai 20, with voltage at 200 kV. The images are 2048 × 2048 in pixel size and 0.4 nm in pixel width. A piece of the images is shown in Fig. 5(b). This sample has a high thickness and tilt-axis shift, which demonstrates the unsuitability of the correlation-based method in such condition.

The third sample is a tilt series of a centriole digitized by a Gaian Camera on TF30, with voltage at 300 kV. It is from sample data provided by IMOD for fiducial marker alignment. The tilt angles of projection images range from +65.0° to −61.0° at 2° intervals. The images are 1024 × 1024 in pixel size and 1.01 nm in pixel width. A piece of the images is shown in Fig. 5(c). This sample provided a comparison between the results of our method and the fiducial marker alignment of IMOD.

3.2. Alignment and reconstruction results

Fig. 6 illustrates the histograms of track length obtained by our method for the three samples. In practice, our method matches 3 adjacent views for each image and then combines the matching pairs to obtain tracks. As shown in Fig. 6, a huge number of tracks are detected in every sample. The number of tracks with a length greater than 3 is more than 5000 for each sample. In mitochondria, the number of tracks with a length greater than 3 is more than 10,000, as shown in Fig. 6(a). In these three samples, the longest tracks cover 101, 72 and 40 images, respectively. All of these values strongly support the effectiveness of our feature-matching and tracking method.

If the rotation matrix R, translation vector t and 3D spatial points are fixed, a snapshot of the projection procedure can be revealed. Fig. 7(a) and 7(b) show the aligned state of the projection procedure using the cross-correlation alignment of IMOD and our method, respectively. The figures are obtained as follows. For each view, the estimated 3D points are fitted to a flat surface in space. For a given view, this surface indicates the exact state of the sample in projection procedure (ignoring the sample thickness). All of the surfaces in different views are superimposed together in a coincident coordinate system. Fig. 7(a) shows that sample translation along the z-axis occurred during the projection procedure, and the displacement of the pitch angle was not even. The rotation axis can hardly be observed in Fig. 7(a). However, the surfaces clearly tilt around a “shaft” in Fig. 7(b). Further analysis of the “shaft” indicated that the intersection of each surface does not coincide; in other words, the pitch angles of the “shaft” have slight differences. Such differences along z-axis are unrecoverable by in-plane translation or rotation of the projection. Although such a direct visualization could not determine the accuracy of an alignment, it really...
reveals the nature of problems. First, as the translation along the $z$-axis showed in Fig. 7(a), we conclude that, using the cross-correlation method, it is difficult to recover an accurate scaling factor of the tilt series. Secondly, because the in-plane rotation and displacement of pitch-angle are indistinguishable by the cross-correlation method, these values are also hard to recover.

The characteristics of the three samples and details of the alignment results are summarized in Table 1. All experiments were performed on a PC with CPU Intel Xeon 3.3 GHz, memory DDR3 8G byte, and system Ubuntu 12.04 64-bit. The performance of our method can be evaluated by the running time and reprojection error, as shown in Table 1. For the mitochondria and caveola samples with images of $2048 \times 2048$, our method can complete the alignment in 40 min. For the centriole sample with images of $1024 \times 1024$, our method can finish the alignment in less than 10 min. The reprojection error is a common measure of alignment accuracy and gives a good indication of the quality of the subsequent reconstruction (Lawrence et al., 2006; Brandt and Ziese, 2006; Baldwin and Penczek, 2005). Here, the reprojection error is calculated as the root mean square (rms) distance between the estimated feature location and the measured feature location, as in the following formula:

$$ E = \sqrt{\frac{1}{M} \sum_{i=1}^{M} |\mathbf{p}_i - \mathbf{p}_i'|^2} $$

where $\mathbf{p}_i$ is the measured points vector and $\mathbf{p}_i'$ is the estimated value. $M$ indicates the total number of estimated projection feature points. As shown in Table 1, the reprojection errors ($e$) of the three

---

**Table 1**

Summary of the statistics for the samples and alignment results.

<table>
<thead>
<tr>
<th></th>
<th>Mitochondria</th>
<th>Caveola</th>
<th>Centriole</th>
</tr>
</thead>
<tbody>
<tr>
<td>Voltage (kV)</td>
<td>200</td>
<td>200</td>
<td>300</td>
</tr>
<tr>
<td>Image size</td>
<td>$2048 \times 2048$</td>
<td>$2048 \times 2048$</td>
<td>$1024 \times 1024$</td>
</tr>
<tr>
<td>No. views</td>
<td>112</td>
<td>119</td>
<td>64</td>
</tr>
<tr>
<td>Pixel width (nm)</td>
<td>0.4</td>
<td>0.4</td>
<td>1.01</td>
</tr>
<tr>
<td>No. chains</td>
<td>29402</td>
<td>13351</td>
<td>8986</td>
</tr>
<tr>
<td>No. features (/image)</td>
<td>1508.06</td>
<td>507.02</td>
<td>569.30</td>
</tr>
<tr>
<td>Mean chain length</td>
<td>5.74</td>
<td>4.44</td>
<td>4.05</td>
</tr>
<tr>
<td>Max chain length</td>
<td>101</td>
<td>72</td>
<td>40</td>
</tr>
<tr>
<td>Running Time (minutes)</td>
<td>37.98</td>
<td>38.32</td>
<td>9.13</td>
</tr>
<tr>
<td>rms error</td>
<td>0.502</td>
<td>0.619</td>
<td>0.469</td>
</tr>
</tbody>
</table>
samples are 0.502, 0.619 and 0.469. The distributions of the reprojection error for every sample are provided for further discussion in Fig. 8. We found that the distributions of the reprojection error are very similar. The distributions are concentrated on 0–2 pixels, while the median value is around 0.4–0.5 (brown line) and the 0.75 quantile is around 0.7–0.8 (red line) for all distributions. Such a positively skewed distribution strongly indicates that our method is robust to outliers, and the reprojection is with subpixel accuracy in total. The rms error of the centriole we got is 0.469, while the reported residual error mean is 0.2–0.5 using the fiducial marker alignment according to the IMOD’s website. Considering the error distribution, the alignment quality is better than the rms error reflected.

In addition to reprojection error, visual clues in the reconstruction are usually used to measure resolution by researchers in practice. Visual clues typically include the edge of the sample, membranes, the shape of gold nano particles and artifacts. Fig. 9 is a comparison of mitochondrial reconstruction based on the cross-correlation alignment of IMOD and our method. These two reconstructions were obtained by the same reconstruction method (weighted back projection, WBP) but different alignment methods. Fig. 9(a) and (b) shows the central x–y sections of two reconstructions; Fig. 9(c) and (d) shows typical y–z sections of two reconstructions. The thickness of the sample is about 150 pixels. This sample thickness is suitable for the cross-correlation alignment method, but our method aligned much better than the cross-correlation method. An in-plane rotation of approximately ±2.5° was detected and corrected by our method. In Fig. 9, we found that the sharpness is similar in both reconstructions, while the reconstruction based on our method shows clearer details of membranes and less artifacts in the y–z section, especially in the selected areas (red rectangle) and the upper and lower borders of the y–z section.

Fig. 10 compares the WBP reconstructions of the caveola based on the cross-correlation alignment of IMOD and our method. Fig. 10(a) and (b) show the corresponding x–y sections of two reconstructions; Fig. 10(c) and (d) shows typical y–z sections of two reconstructions. The thickness of the sample is approximately 400 pixels. Generally, the cross-correlation method cannot obtain desirable alignment at such a thickness, and an in-plane rotation of approximately ±2.5° and an absolute tile axis shift reaching 11°–13° can be detected using our method. Furthermore, in Fig. 10, it is much easier to distinguish biological details in (b) compared to (a). Additionally, the obvious distribution of artifacts in (c) has been significantly reduced in (d).

Fig. 11 illustrates the comparison of centriole WBP reconstruction based on cross-correlation alignment of IMOD, marker alignment and our method. This sample has very good characteristics, including well-distributed fiducial markers, a pre-known gross rotation, sufficient biological details and small thickness. We manually selected about 20 fiducial markers as seeds during the marker alignment. In Fig. 11, the reconstructions based on fiducial marker alignment and our method are very similar in terms of the biological details and marker sharpness. However, it is obvious that the markers carry significant distortion in the top x–y section of the reconstruction based on the cross-correlation alignment. The difference is also observed in the y–z sections.

### 3.3. Quantitative analysis of alignment quality

The accuracy of alignment could be directly reflected by the quality of reconstruction tomography. However, estimation of the reconstruction quality in ET still remains an open problem (Leis et al., 2009; Henderson et al., 2012). Thus, in our quantitative analysis, two different criteria are employed to evaluate the improvement of alignment accuracy. The first one is the normalized correlation coefficient between the input tilt series and the corresponding reprojection series of the reconstruction. The coefficient could give a comparison of alignment quality for each image as a function of the tilt angle but not the quality of the whole reconstruction directly. The second one is the Fourier Shell Correlation splitting the tilt series into two halves (FSC_{y/0}), which is a resolution estimation criteria proposed by (Cardone et al., 2005) and has been popularly used in previous works (Fernández, 2012). The FSC_{y/0} is used here to show the improvement of alignment quality in quantitative terms, but does not necessarily reflect the actual resolution of the final tomograms.

Fig. 12 shows the variation of the coefficient values with the tilt angle, where coarse alignment refers to the cross-correlation method provided by IMOD and fine alignment refers to the fiducial marker alignment in IMOD. For every sample, we have already built the tomograms by WBP in the previous section. Then, the reconstructions obtained by our method and by cross-correlation are reprojected according to the tilt parameters, thus calculating the corresponding series of reprojections. Finally, the normalized correlation coefficient between the calculated reprojection and original projection in each tilt angle are computed. As shown in Fig. 12(a) and (b), there is a big gap between the curves that were obtained, which indicates that our method outperforms cross-correlation. Furthermore, in Fig. 12(a), we can find a significant vibration around the tilt number of 55 due to misalignment in the result of cross-correlation, while there is a remarkable improvement in the result of ours. Fig. 12(c) gives a further comparison of the results between our method and the fiducial marker alignment. Here, the difference among the three methods is not so significant, but still we should note that both the result of our method and that of marker alignment are better than the one of cross-correlation. The curve of our method is not so good as the curve of marker alignment in the front segment while it is a little better in the middle segment and almost coincident with the one of marker alignment after the tilt number of 40.

Though normalized correlation coefficient could give a comparison of alignment quality for each image at different tilt angle, it could not provide the comparison of quality improvement of the
reconstruction directly. FSC_{e,o} is adopted to estimate the resolution of reconstructions (FSC value is calculated by Bsoft (Heymann and Belnap, 2007)). According to Cardone et al. (2005)’s investigation, though the FSC method is directly adopted from Single-Particle Analysis, it is in agreement with the other criteria proposed in Cardone’s work. One point we should note is that Cardone’s work also reports that FSC_{e,o} underestimates the actual resolution. Nevertheless, though FSC_{e,o} give the resolution value in an underestimated way, FSC_{e,o} could reflect the trends, and has been popularly used in the scope, so it is used as the resolution criteria here.

To avoid the influence of edges, we choose the central volume (Table 2) of the reconstructions for estimation. As illustrated in Table 2, the resolution obtained by our method is 18.5 Å for mitochondria, 26.6 Å for caveola and 50.5 Å for centriole, better than the corresponding reconstruction obtained by cross-correlation alignment of IMOD. The details of the FSC_{e,o} curve for every specimen are shown in Fig. 13. Note that the FSC_{e,o} curves of our method and fiducial marker alignment are almost coincident for the centriole.

It should be noted that the distribution of resolution is not even. That is, the background or noise averages the FSC_{e,o} value, which means that the global FSC_{e,o} underestimates the effect of our method. For example, if we choose the area of 512 × 512 around the red rectangle near the center in the caveola, as shown in Fig. 10, the FSC_{e,o} (0.5) value of the selected area obtained by our method is 26.9 Å, while the corresponding value obtained by cross-correlation alignment is 29.5 Å. The FSC_{e,o} curves of the selected area are illustrated in Fig. 14, where there is a large gap between the two curves.

However, such a comparison is not sustainable, as some areas can certainly be found where our method is not better than the

![Fig. 9](image9.png)

(a) The central x–y section of the reconstruction of mitochondria aligned by IMOD. (b) A similar x–y section of mitochondria aligned by our method. (c) A typical y–z section of the reconstruction of mitochondria aligned by IMOD. (d) A similar y–z section of mitochondria aligned by our method. The structure of the mitochondria is clearly visible in all reconstructions, but the reconstruction with our method shows clearer details and considerably less background noise.

![Fig. 10](image10.png)

(a) The central x–y section of the reconstruction of the caveola aligned by IMOD. (b) A similar x–y section of the caveola aligned by our method. (c) A typical y–z section of the reconstruction of the caveola aligned by IMOD. (d) A similar y–z section of the caveola aligned by our method. The sharpness of details is similar in both reconstructions, but the background noise is considerably smaller in the reconstruction by our method, and the artifacts are notably reduced in the y–z section.
Fig. 11. Comparison of the reconstruction of the centriole after cross-correlation alignment (left), marker alignment (middle) and our method (right). (a), (b), and (c) are similar top \(x-y\) sections. (d), (e), and (f) are similar middle \(x-y\) sections. (g), (h), and (i) are similar \(y-z\) sections. Note that the results of our method and those of the marker alignment have similar sharpness, while the result of cross-correlation alignment exhibits obvious distortions in the top \(x-y\) section.

Fig. 12. Normalized correlation coefficient. (a) Mitochondria. (b) Caveola. (c) Centriole.

Table 2

<table>
<thead>
<tr>
<th></th>
<th>Mitochondria</th>
<th>Caveola</th>
<th>Centriole</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pixel width (nm)</td>
<td>0.4</td>
<td>0.4</td>
<td>1.01</td>
</tr>
<tr>
<td>Central volume selected (pixel)</td>
<td>(1800 \times 1800 \times 150)</td>
<td>(1024 \times 1024 \times 388)</td>
<td>(800 \times 800 \times 125)</td>
</tr>
<tr>
<td>FSC(_{1/2}(0.5)) of cross-correlation ((\AA))</td>
<td>18.9</td>
<td>27.6</td>
<td>51.3</td>
</tr>
<tr>
<td>FSC(_{1/2}(0.5)) of our method ((\AA))</td>
<td>18.5</td>
<td>26.6</td>
<td>50.5</td>
</tr>
<tr>
<td>FSC(_{1/2}(0.5)) of fine ali. ((\AA))(^a)</td>
<td>—</td>
<td>—</td>
<td>50.4</td>
</tr>
</tbody>
</table>

\(^a\) Fine ali. refers to the fiducial marker alignment of IMOD.
cross-correlation alignment. For a comprehensive evaluation of the effectiveness of our method, we designed an experiment using sublocal resolution. First, we divided the central volumes labeled in Table 2 into several subvolumes. Adjacent subvolumes overlap each other, as shown in Fig. 15. We then calculated the \( FSC_{e=0} \) value for each subvolume.

For mitochondria, we divided the central volume of \( 1800 \times 1800 \times 150 \) into \( 2025(45 \times 45) \) subvolumes, with each volume \( 360 \times 360 \times 150 \). Fig. 16 shows the \( FSC_{e=0} \) values of each subvolume in a pseudo-color map. Compared to the cross-correlation alignment, our method did much better on the top left, top right and bottom right areas. The results are also coincident with the visual estimation, as in Fig. 9, where the selected areas in the red rectangle show clear differences.

For the caveola, we divided the central volume of \( 1024 \times 1024 \times 388 \) into \( 900 (30 \times 30) \) subvolumes, with each volume \( 341 \times 341 \times 388 \). Fig. 17 demonstrates the \( FSC_{e=0} \) values of each subvolume in a pseudo-color map. There is a significant improvement in the right central area, where it is red for cross-correlation alignment compared to light green for our method, as shown in Fig. 17.

For the centriole, we divided the central volume of \( 800 \times 800 \times 125 \) into \( 900 (30 \times 30) \) subvolumes, with each volume \( 250 \times 250 \times 125 \). Fig. 18 shows the \( FSC_{e=0} \) values of each subvolume in a pseudo-color map. The result of our method (Fig. 18(c)) is almost coincident with that of the fiducial alignment (Fig. 18(b)), and is better than that of cross-correlation alignment (Fig. 11(a)) in the top left area. Fig. 11 shows that the left center centriole structure (around the red rectangle in Fig. 11(a)) is blurrier than the corresponding areas in (b) and (c).

In conclusion, the analysis of subvolumes is complementary to the global \( FSC_{e=0} \) analysis of reconstruction. Although the analysis of subvolumes is based on the local region, it gives us a comprehensive impression of the resolution distribution in a tomographic reconstruction as a global level. Both the results of subvolume \( FSC_{e=0} \) and global \( FSC_{e=0} \) are coincident with the analysis done by
the normalized correlation coefficient and previous visual clues. All the analysis supports the effectiveness of our method.

4. Discussion and conclusion

In this paper, we propose a new marker-free alignment method based on scale-invariant features. Our focus is to utilize the abundant information of biological structures in micrographs to optimize the projection parameters. The experimental results indicate that our method guarantees accurate alignment that is comparable to fiducial marker alignment.

Although our method is quite effective in processing marker-free tilt series, there still remain some aspects for further investigation.

4.1. Assumptions, limitations and potential applicability

The image intensity at each point of a micrograph is related to the exponential of a line integral along a specific trajectory through the object (as the fundamental of ray transform (Popov, 2001; Lawrence et al., 2006)). Thus, the performance of our method has to do with the intrinsic biological structure in the sample. To optimize the projection parameters, our method needs a minimum number of measured features. Otherwise, this method cannot achieve accurate alignment without sufficient features. Furthermore, a feature in the projection does not always correspond to a feature in the object. For example, in some cases, the integral of the object density may generate “false features” in micrograph if multiple object features superposed along the projection trajectory. Thus, the performance of our method is limited by the thickness of samples, which affects the localizing accuracy of features. Our method can benefit from high-contrast structures in the specimen since high-contrast structures have a major contribution to the image intensity in a micrograph to avoid feature superposition. Sample’s thickness also hampers the performance in the aspect that the increasing of thickness raises the risk of deformation, which cannot be fixed without further knowledge.

Some features in the projection may change considerably in different tilts due to the effect of linear integral. However, the change of features, or even the loss of some is not a big problem. Our approach can adjust the projection parameters well if sufficient information is extracted, because an incremental bundle adjustment strategy has been introduced to our approach, as described in the Section 2.4. The bundle adjustment strategy can improve our method so as to estimate the projection parameters by “local information” of relative short tracks.

SNR of the micrographs is another potential limitation of our method. Our method may probably have no effect on cryo-ET tilt series, which are characterized by a particularly low SNR (a similar topic has also been discussed by Kaynig et al. (2010)). In the field of cryo-ET, marker-free alignment is still a challenge (Castañó-Díez et al., 2010). CTF (Contrast transfer function) may also affect the performance of our method by degenerating the localizing accuracy of features. We do not consider the effect of CTF because in general it is only important for high resolution studies by cryo-ET and subtomogram averaging (Fernández et al., 2006). In the future, we will go on with this investigation. However, these limitations do not detract from the good behavior of our method in other widespread modalities of electron tomography where the contrast is better.

Our method could be easily extended into multi-axis alignment or other versions of tomography, soft X-ray tomography, etc. Our method also has potential applicability in materials science, where the embedment of fiducial markers is often impossible (Fernández, 2013).
4.2. Process acceleration and parameter model extension

As illustrated previously, the running time of our method for a typical tilt series of $2048 \times 2048$ is approximately 40 min. Compared with the execution time of the cross-correlation method (about 5 min), it is quite time-consuming. In our method, feature extraction and parameter optimization take a long processing time. Fortunately, our method has a good intrinsic property for parallelization, thus, MPI, CUDA and even multi-core parallel techniques (Fernández, 2008) could be used to accelerate our alignment method.

Another issue we should point out is the uneven distortion of views. Some irregular distortion may be caused by sample deformation, while Lawrence et al. (2006); Phan et al. (2009) introduced the correction of non-linear lens distortions into large field electron tomography and have achieved impressive results. A more parameterized projection model is required to recover projection distortion. For example, a quadratic model and a cubic model are presented in Lawrence et al. (2006). However, the more parameters a model has, more severe the tracking is needed. So far, all existing non-linear parameterized projection models are (partly) based on fiducial alignment. Compared to previous feature-based alignment methods, our method has a potential extension to recover partial distortion of which is the rich track information. Further work to recover projection distortion in marker-free conditions is being performed.

5. Implementation

A software compatible with IMOD based on our method, atoma-align, is under development. Interested users can download the beta version at our site http://ear.ict.ac.cn/?page_id=34.

Acknowledgments

This work is supported by grants National Natural Science Foundation for China (61232001, 61202210 and 60921002) and grants to F.S. from the Chinese Ministry of Science and Technology (2014CB910700). J.-J.Fernández thanks the support from Spanish grant TIN2012–37483. The authors would like to thank Phd student Fan Xu, Jingrong Zhang for their helps about this work.

Appendix A. Sparse implementation of adjustment

The L-M algorithm (Kanzow et al., 2004) is employed to solve the non-linear least squares. Deducing from the cost function, we solve the following equations iteratively:

$$\Phi = (f^T + \mu I)^{-1/2}fr$$

where $\Phi = (\Theta, X)$, $\Theta$ indicates projection parameters, $X$ is the spatial points and $r$ is residual (i.e., difference between measured image points and estimated value over the projection process). $J = \partial r / \partial \Theta$. $J$ is the $M \times N$ Jacobian matrix, $M$ is the number of measured image point and $N = n_\Theta + n_X$, which is the sum of the number of projection parameter numbers and spatial points.

To solve Eq. (A.1) directly is possible but ineffective. However, the abundant zero value in $J$ could be utilized, because some spatial points only have a few projection points. Consider the structure of $J$.

$$J^TJ = \begin{pmatrix}
\frac{\partial J}{\partial \Theta_1} & \frac{\partial J}{\partial \Theta_2} & \cdots & \frac{\partial J}{\partial \Theta_p} \\
\frac{\partial J}{\partial X_1} & \frac{\partial J}{\partial X_2} & \cdots & \frac{\partial J}{\partial X_Q}
\end{pmatrix}
\begin{pmatrix}
\frac{\partial J}{\partial \Theta_1} & \frac{\partial J}{\partial \Theta_2} & \cdots & \frac{\partial J}{\partial \Theta_p} \\
\frac{\partial J}{\partial X_1} & \frac{\partial J}{\partial X_2} & \cdots & \frac{\partial J}{\partial X_Q}
\end{pmatrix}^T
$$

We found that $C_{\Theta\Theta}^{-1}$ and $C_{X\Theta}^{-1}$ are both diagonal matrices. By substituting expressions from Eq. (A.3) to Eq. (A.5), Eq. (A.1) can be rewritten as the following linear equation:

$$\begin{pmatrix} A & B \\ B^T & C \end{pmatrix}
\begin{pmatrix} \Theta \\ X \end{pmatrix} =
\begin{pmatrix} e_\Theta \\ e_X \end{pmatrix}$$

where $A = C_{\Theta\Theta}^{-1} + \mu I$, $C = C_{X\Theta}^{-1} + \mu I$. $B = C_{\Theta X} \hat{r}$ and $e_X = \frac{\partial f}{\partial \Theta}. r$. A further term of Eq. (A.6) is:

$$\begin{pmatrix} A - BC^{-1}B^T \\ B^T \end{pmatrix}
\begin{pmatrix} \Theta \\ X \end{pmatrix} =
\begin{pmatrix} e_\Theta - BC^{-1}e_X \\ e_X \end{pmatrix}$$

Therefore, $\Theta$ can be solved from the top half of Eq. (A.7), and $X$ can be solved from the bottom half. By employing the sparse linear problem, we reduce the scale of the problem.

References


