Bioinformatics, 34(17), 2018, 3013–3021
doi: 10.1093/bioinformatics/bty210
Advance Access Publication Date: 19 April 2018
Original Paper

OXFORD

Bioimage informatics

Comparative analysis of tissue reconstruction algorithms for 3D histology

Kimmo Kartasalo[1,2,3], Leena Latonen[1,3,4], Jorma Vihinen[5], Tapio Visakorpi[1,3,4], Matti Nykter[1,2,3] and Pekka Ruusuvuori[1,3,6,*]

[1]Faculty of Medicine and Life Sciences, University of Tampere, Tampere 33014, Finland, [2]Faculty of Biomedical Sciences and Engineering, Tampere University of Technology, Tampere 33101, Finland, [3]BioMediTech Institute, Tampere 33014, Finland, [4]Fimlab Laboratories, Tampere University Hospital, Tampere 33101, Finland, [5]Faculty of Engineering Sciences and [6]Faculty of Computing and Electrical Engineering, Tampere University of Technology, Tampere 33101, Finland

*To whom correspondence should be addressed.
Associate Editor: Robert Murphy

Received on November 29, 2017; revised on March 1, 2018; editorial decision on March 24, 2018; accepted on April 18, 2018

Abstract

Motivation: Digital pathology enables new approaches that expand beyond storage, visualization or analysis of histological samples in digital format. One novel opportunity is 3D histology, where a three-dimensional reconstruction of the sample is formed computationally based on serial tissue sections. This allows examining tissue architecture in 3D, for example, for diagnostic purposes. Importantly, 3D histology enables joint mapping of cellular morphology with spatially resolved omics data in the true 3D context of the tissue at microscopic resolution. Several algorithms have been proposed for the reconstruction task, but a quantitative comparison of their accuracy is lacking.
Results: We developed a benchmarking framework to evaluate the accuracy of several free and commercial 3D reconstruction methods using two whole slide image datasets. The results provide a solid basis for further development and application of 3D histology algorithms and indicate that methods capable of compensating for local tissue deformation are superior to simpler approaches.
Availability and implementation: Code: https://github.com/BioimageInformaticsTampere/RegBenchmark. Whole slide image datasets: http://urn.fi/urn: nbn: fi: csc-kata20170705131652639702.
Contact: pekka.ruusuvuori@tut.fi
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Digitalization of pathology has been accelerated by improvements in technology allowing acquisition of whole slide images (WSI) (Ghaznavi *et al.*, 2013; Griffin and Treanor, 2017). Besides computer-aided facilitation of pathologists' tasks, digital pathology can enable new approaches like 3D histology, where three-dimensional reconstructions of samples are formed *in silico* based on serial sections (Magee *et al.*, 2015; Roberts *et al.*, 2012). While other techniques allow imaging directly in 3D, they are currently incapable of matching the subcellular resolution and throughput of whole slide imaging. Examples of potential applications include construction of data-driven computer models and improved diagnostics of diseases associated with changes in the 3D microarchitecture of tissue. Moreover, 3D histology is compatible with established histopathological interpretation techniques and biochemical assays such as immunohistochemistry or *in situ* hybridization. This raises interesting prospects in view of recent advances in spatially resolved omics (Mignardi *et al.*, 2017; Ståhl *et al.*, 2016). Pairing imaging with genomic, epigenomic, transcriptomic and proteomic data in the spatial context of tissue holds great promise for pathology and other fields (Koos *et al.*, 2015). Taking a step further, this could be performed in 3D to truly probe the relationships between structural and functional features as well as the heterogeneity and interplay between different cell types in tumors, and significant projects are

now pursuing these goals (Ledford, 2017; Rusk, 2016). These kind of approaches have already led to the creation of brain atlases (Amunts *et al.*, 2013; Johnson *et al.*, 2010; Lein *et al.*, 2007). Such high-dimensional data also represent an exciting challenge for new ways of scientific visualization based e.g. on virtual reality techniques (Calì *et al.*, 2016; Ledford, 2017; Theart *et al.*, 2017).

Despite earlier computational and image acquisition bottlenecks (Roberts *et al.*, 2012), several algorithmic 3D histology solutions were already proposed before the recent developments in digital pathology (Ju *et al.*, 2006; Wang *et al.*, 2015). The key methodological problem is how to accurately register a sequence of 2D images to produce a 3D volume. Simply stacking the images does not result in a coherent volume due to differences between the relative locations and rotation angles of the sections and tissue deformations introduced during embedding and sectioning (Gibson *et al.*, 2013). Algorithms for image registration (Sotiras *et al.*, 2013) constitute the methodological basis of 3D histology. These algorithms are used to sequentially register each image with its neighbors to bring the entire series into alignment (Magee *et al.*, 2015; Wang *et al.*, 2015). Registration is accomplished by estimating transformations relating the images. Rigid transformations only allow translation and rotation of the entire image, while affine transformations are additionally able to model anisotropic scaling. Locally varying transformations, also called elastic models, can compensate for deformations on a local scale. Considering several nearby sections together (Saalfeld *et al.*, 2012) or applying regularization may be needed to obtain smooth, continuous 3D volumes (Casero *et al.*, 2017; Cifor *et al.*, 2011; Gaffling *et al.*, 2015; Ju *et al.*, 2006). After estimating the transformations, they need to be applied to the images via interpolation, which is possibly followed by postprocessing such as 3D visualization. Our focus is on the reconstruction step, which is usually the most difficult and crucial part of the image processing chain. Numerous approaches have been reported, relying on manual alignment (Onozato *et al.*, 2012; Paish *et al.*, 2009), semi-automatic methods using artificial landmarks (Hughes *et al.*, 2013; Rojas *et al.*, 2015) and automated algorithms (Arganda-Carreras *et al.*, 2010; Braumann *et al.*, 2005; Casero *et al.*, 2017; Cifor *et al.*, 2011; Ju *et al.*, 2006; Magee *et al.*, 2015; Saalfeld *et al.*, 2012; Song *et al.*, 2013; Stille *et al.*, 2013; Xu *et al.*, 2015).

Despite the widely acknowledged need for objective assessment of algorithms (Meijering *et al.*, 2016), an evaluation of modern computational methodology for 3D histology is lacking. Moreover, the common practice of relying only on visual inspections or a single indirect metric is insufficient (Rohlfing, 2012). The previous comparison of algorithms was published a decade ago and only included three basic approaches (Beare *et al.*, 2008). We have previously demonstrated a framework (Kartasalo *et al.*, 2016) based on a panel of indirect metrics and manually annotated landmarks allowing direct quantification of reconstruction accuracy (Rohlfing, 2012). In this study, we applied an extended version of the framework (see Fig. 1) to address the problem of comparing algorithms for 3D histology. As the basis of our evaluation, we used two WSI datasets representing two different tissue types. One obstacle complicating both the application and fair comparison of most algorithms is sensitivity to various settings or hyperparameters, which typically have to be selected by the user based on rules of thumb and tuned via trial and error. Encouraged by their recent application in the context of digital pathology, we employed automated hyperparameter selection methods to adjust tunable parameters (Shahriari *et al.*, 2016; Teodoro *et al.*, 2017).

As a baseline, we evaluated three basic methods: a least-squares fit to landmarks (LS), an optimization-based approach (OPT) and a

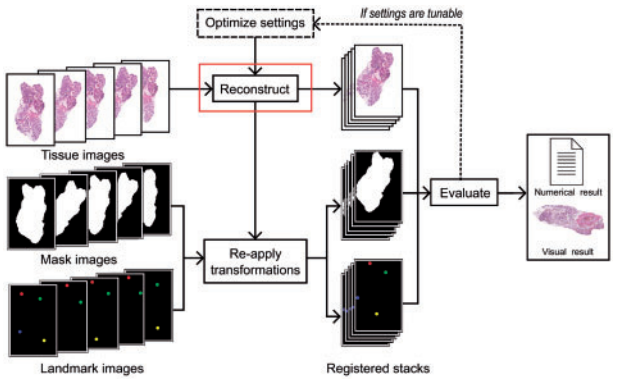

Fig. 1. Evaluation framework. A series of tissue images is input to a reconstruction method for registration. The transformations estimated by the method are re-applied to masks defining the tissue region and images containing landmarks. The registered tissue, mask and landmark images are used to evaluate reconstruction accuracy based on numerical metrics and visual examination. Moreover, tunable settings can be optimized. (Color version of this figure is available at *Bioinformatics* online.)

method based on the Scale Invariant Feature Transform (SIFT) (Lowe, 2004). More advanced methods included the Fiji/ImageJ (Schindelin *et al.*, 2012; Schneider *et al.*, 2012) plugins HyperStackReg (HSR), which is an extension of StackReg (Thevenaz *et al.*, 1998), RegisterVirtualStackSlices (RVSS), which is based on bUnwarpJ (Arganda-Carreras *et al.*, 2006), and ElasticStackAlignment (ESA) (Saalfeld *et al.*, 2012), which is part of the TrakEM2 package (Cardona *et al.*, 2012). In addition, we evaluated two commercial tools: Medical Image Manager (MIM) (HeteroGenius Ltd, Leeds, UK) and Voloom (microDimensions GmbH, Munich, Germany). While LS, OPT, SIFT and HSR are based on global transformations, RVSS, ESA, MIM and Voloom use elastic models which make it possible to account for local tissue deformations. For a summary of the evaluated tools, see Supplementary Table S1.

2 Materials and methods

2.1 Data collection and preprocessing

A murine prostate and a liver were fixed in PAXgene™ (PreAnalytiX GmbH, Hombrechtikon, Switzerland) and formalin, respectively, embedded in paraffin, and cut into serial 5 μm sections. The liver was processed with a laser prior to embedding in order to introduce artificial landmarks into the otherwise homogeneous tissue. Four holes were successfully introduced into the sample. The sections were hematoxylin-eosin (HE) stained and scanned at 20× (pixel size 0.46 μm) to obtain 260 (prostate) and 47 (liver) RGB images. The images were processed in MATLAB R2016b (The MathWorks Inc., Natick, MA, USA) to segment tissue from background and store the results as binary masks.

A total of 2448 landmarks were manually annotated. In the prostatic tissue, four corresponding points preferably at the centers of bisected nuclei were selected by two observers from each pair of adjacent sections. For the liver, the four holes in each image were marked by the same two observers. Most of the evaluated methods do not allow direct application of transformations to coordinates but support re-applying them to another stack of images. Therefore, we stored the landmarks as images with four disks placed at the landmark locations, each consisting of red, green, blue or yellow pixels. Color is invariant to the applied transformations, allowing

post-registration detection of the disks. The tissue, mask and landmark images were downsampled to different resolutions and stored as TIF. See Supplementary Methods for details.

2.2 Evaluation of reconstruction accuracy

2.2.1 Target registration error

Pairwise target registration error (TRE) (Fitzpatrick *et al.*, 1998), a direct measure of registration accuracy (Rohlfing, 2012), was quantified for each pair of adjacent sections. From the landmark images, we detected each landmark based on the colors of the disks and obtained their coordinates as the centroids of the detected pixels. For N pairs of sections, TRE was measured for each point ($j = \{1, 2, 3, 4\}$) and section pair ($i = \{1, 2, \ldots, N\}$) as:

$$TRE_{j,i} = \|\boldsymbol{X}_{j,i} - \boldsymbol{X}_{j,i+1}\| \tag{1}$$

that is, the Euclidean distance between the location $\boldsymbol{X_{j,i}}$ of point j on the section i and the location of the corresponding point on section $i+1$.

2.2.2 Accumulated error

Accumulated target registration error (ATRE) was calculated to quantify distortion accumulated through the stack, referred to as 'the banana problem' (Malandain *et al.*, 2004) or 'the shear effect' (Hughes *et al.*, 2013). Each landmark of the prostate dataset is only present on two consecutive sections and pairwise errors on different sections should thus be independent of each other. However, in the presence of accumulated errors, the error vectors on nearby sections are correlated (Beare *et al.*, 2008). We quantified this effect by treating the displacement of each landmark ($j = \{1, 2, 3, 4\}$) for each pair of sections ($i = \{1, 2, \ldots, N\}$) in vector form as $\boldsymbol{X}_{j,i} - \boldsymbol{X}_{j,i+1}$ and averaging the four vectors to obtain the mean displacement of each entire section. We then computed the cumulative sum of these mean vectors, proceeding from section 1 to section N. For section k, ATRE was defined as the Euclidean norm of the cumulative displacement vector:

$$ATRE_k = \left\| \sum_{i=1}^{k} \sum_{j=1}^{4} \frac{\boldsymbol{X}_{j,i} - \boldsymbol{X}_{j,i+1}}{4} \right\| \tag{2}$$

For the liver, a more direct quantification of ATRE was possible due to the landmarks extending through the sample. Ideally, the landmarks should lie on four parallel lines. In practice, parallelism could be violated due to slight movement of the sample between repeated applications of the laser. In a distorted volume, the landmarks deviate from the linear trajectories when proceeding through the stack. To measure this, we fitted a line in 3D to each of the four series of landmarks, minimizing mean squared error on the image plane. ATRE was then quantified for section i and landmark j as the Euclidean distance between the location of the landmark $\boldsymbol{X_{j,i}}$ and that of the fitted line $\boldsymbol{Y_{j,i}}$, on the image plane:

$$ATRE_{j,i} = \|\boldsymbol{X}_{j,i} - \boldsymbol{Y}_{j,i}\| \tag{3}$$

2.2.3 Tissue shrinkage and overlap

As certain reconstruction methods tend to shrink the tissue, relative change in tissue area (ΔA-%) was computed based on the tissue masks for each section. Overlap was quantified based on the masks for each section pair using the Jaccard index (Rohlfing, 2012). The Jaccard index can be considered a quality measure for pixel-wise metrics, as computing them for a pair of sections with little overlap can provide misleading results. Let A denote the set of tissue pixels of section i and B the set of tissue pixels of section $i+1$. The Jaccard index is defined as:

$$Jaccard_i = \frac{|A \cap B|}{|A \cup B|} \tag{4}$$

2.2.4 Pixel-wise similarity

For each section pair, we evaluated the similarity of corresponding pixels. After conversion to grayscale we computed the following measures: root mean squared error (RMSE), normalized cross correlation (NCC), mutual information (MI) and normalized mutual information (NMI) (Studholme *et al.*, 1999). Only the set of overlapping tissue pixels $A \cap B$ was considered. These indirect metrics provide information from the entire tissue area and complement the TRE evaluation.

2.2.5 Reconstruction smoothness

We quantified the smoothness of the reconstruction using contrast f_2 and correlation f_3 based on gray-level co-occurrence matrices (GLCMs) (Cifor *et al.*, 2011; Gaffling *et al.*, 2015; Haralick and Shanmugam, 1973). Low contrast and high correlation indicate a smooth reconstruction. We formed the GLCM for each pair of grayscale images based on pixels $A \cap B$ and summed them to obtain a single GLCM for the whole volume.

2.3 3D reconstruction

- LS: Least-squares fitting of an affine transformation to the landmarks was implemented in MATLAB R2016b. The result is in principle unaffected by error accumulation (Xu *et al.*, 2015).
- OPT: Optimization-based reconstruction implemented in MATLAB R2016b was used to estimate pairwise affine transformations by minimizing the value of pixel-wise MSE.
- SIFT: Feature-based reconstruction was performed by computing SIFT keypoints (Lowe, 2004) for each image pair, establishing putative matches and robustly fitting an affine transformation to the point pairs (Fischler and Bolles, 1981). We used the RegisterVirtualStackSlices (Arganda-Carreras *et al.*, 2006) implementation in Fiji, also used as an initial step in RVSS and ESA.
- HSR: HyperStackReg v. 5 (Ved P. Sharma, Albert Einstein College, https://sites.google.com/site/vedsharma/imagej-plugins-macros/hyperstackreg) was run in Fiji to perform reconstruction using affine transformations.
- RVSS: Elastic reconstruction based on the bUnwarpJ algorithm, which is a combination of SIFT and optimization based methods, was applied using the RegisterVirtualStackSlices plugin in Fiji.
- ESA: The algorithm implemented in the ElasticStackAlignment plugin (Saalfeld *et al.*, 2012) was run via the TrakEM2 package (Cardona *et al.*, 2012) in Fiji to perform elastic reconstruction based on a combination of SIFT and optimization methods.
- MIM: Medical Image Manager, trial v. 0.94, was applied using images subsampled by a factor of 4 (magnification of 5×) as input. Sections 130 and 24 were used as references for the prostate and liver, respectively. We varied the initial magnification (0.3125×, 0.625×, 1.25× or 2.5×) and the number of non-rigid levels (1, 2, 3 or 4), thus modifying the image resolution used.
- Voloom: Trial v. 2.7.1 was used for elastic 3D reconstruction.

Fiji (Schindelin *et al.*, 2012; Schneider *et al.*, 2012) (v. 1.51h) plugins were run via ImageJ-MATLAB interface (v. 0.7.1) (Hiner *et al.*, 2016). Transformations were re-applied to the mask and landmark

images. Output was saved as TIF. See Supplementary Methods for details.

2.4 Parameter optimization

In the case of MIM, which had to be operated interactively, we evaluated each combination of tunable values by a parameter sweep. Tunable parameters of the other methods were optimized via Bayesian optimization (Shahriari *et al.*, 2016; Snoek *et al.*, 2012), which is well-suited for such problems, where the objective function is computationally expensive to evaluate, nonconvex, multimodal, and typically has low to moderate dimensionality. Bayesian optimization has been shown to perform favorably in comparison to other global optimization algorithms on benchmarking functions (Jones, 2001) as well as on real WSI data (Teodoro *et al.*, 2017). We used MATLAB's *bayesopt* implementation (https://www.mathworks.com/help/stats/bayesian-optimization-algorithm.html) with mean pairwise TRE as the objective function. We utilized a Gaussian process model of the objective function and an automatic relevance determination (ARD) Matérn 5/2 kernel (Snoek *et al.*, 2012) with 'expected-improvement-plus' as the acquisition function (Bull, 2011). Reconstructions with output image dimensions over fivefold compared to the input due to extreme error accumulation were considered failures. The number of variables to optimize was 2 (OPT), 4 (SIFT), 7 (RVSS) or 15 (ESA). We first optimized SIFT alone and used the optimal values for the SIFT step of RVSS and ESA. See Supplementary Table S1 for descriptions of the parameters. The number of seed points was set to twice the number of variables. We ran 30 iterations for OPT due to its simple objective function (Kartasalo *et al.*, 2016) and 100 iterations for the other tools. We used the prostate images subsampled by factors of 8 and 16, except for ESA, for which optimization was only feasible using the factor 16. Parameters optimized for ESA using the lower resolution were scaled to be used with the high resolution images. Computations were run on a workstation with Intel Xeon E5-1660 v3 3 GHz and 64 GB of RAM (low resolution) and a cluster node with Intel Xeon E5-2680 v3 2.5 GHz and 128 GB of RAM (high resolution).

3 Results

3.1 Effect of image resolution on evaluation metrics

First, we analyzed whether our metrics depend on image resolution (see Supplementary Results). TRE, ATRE, Jaccard and ΔA-% are essentially invariant to image resolution. They can be compared across different datasets and resolutions, as long as the accumulation of interpolation errors is avoided. RMSE, NCC, MI, NMI, f_2 and f_3 depend both on resolution and image content, and these metrics should thus only be compared within the same dataset and resolution. In all following analyses, we used images subsampled to pixel sizes of 7.36 and 3.68 μm, referred to as low and high resolution, respectively. The pixel sizes are close to the 5 μm section spacing and metrics computed from these images are not distorted by interpolation errors. Furthermore, we will only present RMSE as a measure of pixelwise similarity and f_2 as a measure of reconstruction smoothness due to their strong correlations with NCC, MI, NMI and f_3 (see Supplementary Table S1 for details).

3.2 Automated parameter tuning

Of the evaluated methods, LS, HSR and Voloom do not have tunable parameters. For OPT, SIFT, RVSS, ESA and MIM, we tuned the parameters automatically, minimizing the mean TRE computed for the prostate dataset. Parameter optimization took approximately 1500 hours in total to compute, producing 23 terabytes of data.

The optimization mostly converged close to the final solution in a handful of iterations (see Supplementary Results). By inspecting the variation in mean TRE values obtained during the process it is possible to reach a semi-quantitative view of the sensitivity of each method towards parameter adjustments. OPT and SIFT produced similar results for most parameter combinations while ESA, MIM and especially RVSS exhibited more sensitivity to parameter tuning.

We evaluated possible connections between accuracy and computation time, which might require the user to make a trade-off when selecting parameters (see Supplementary Results). The time taken by OPT varied only by a few minutes, except for the single inaccurate solutions where the parameters have not allowed proper convergence of the algorithm. For SIFT, there were no signs of a connection between accuracy and computation time. The differences in computation time between the fastest and slowest iterations of RVSS were roughly twofold and the fastest iterations were generally the ones with the highest error, indicating that minimizing the computation time of RVSS would sacrifice accuracy. In the case of ESA, the effect of parameter tuning was dramatic, leading to variation from approximately 12 min to more than 41 h. However, any clear relationship between computation time and accuracy was not observed.

3.3 Comparison of algorithms based on the prostate dataset

Results for the prostate dataset are listed in Table 1. The TRE values of LS based on landmarks by the two observers (LS1 and LS2) establish a baseline of accuracy. The case where the same landmarks were used for reconstruction and for calculating errors (LS1) is an optimistic estimate, representing the best accuracy reachable using an affine model. The errors calculated based on landmarks not used for reconstruction (LS2) represent a more realistic estimate of the accuracy of LS, serving as a cross-validation experiment between the two observers. The discrepancy between the optimistic and cross-validation results indicates that the LS solutions represent overfitting to the landmarks. Therefore, any methods with accuracy approaching LS can be regarded as highly accurate, since the other methods are not provided with any information concerning the landmarks. The systematic difference between TRE and ATRE calculated based on the two sets of landmarks (see Supplementary Table S1) is due to the fact that the two observers were free to select different landmarks and the error is generally not constant over the entire tissue section. However, using either set of landmarks leads to the same conclusions regarding the relative accuracy of the methods, confirmed by linear correlation coefficients of approximately 0.999 for mean TRE, 0.995 for maximum TRE, 0.888 for mean ATRE and 0.901 for maximum ATRE between the two sets of landmarks for the low resolution reconstructions. This also holds for the high resolution with corresponding values of 0.999, 0.986, 0.894 and 0.922. This indicates that even though four landmarks per section pair represent a relatively sparse sampling of the entire tissue section area, this number of landmarks is sufficient for reliable error estimation.

All methods benefited from parameter tuning on both image resolutions based on most of the metrics, using either set of landmarks for evaluation (see Table 1 and Supplementary Results). Of the top three methods, MIM and RVSS obtained better accuracy using high resolution images and ESA worked better on the low resolution images. ESA and MIM reached similar mean TRE values, slightly better than RVSS and approaching or exceeding the accuracy of LS.

Table 1. Evaluation results for the prostate data at low (top) and high resolution (bottom)

Prostate, low resolution

Algorithm	TRE1 μ	TRE1 max	TRE1 σ	ATRE1 μ	ATRE1 max	ATRE1 σ	RMSE μ	RMSE σ	Jaccard μ	Jaccard σ	Contrast f_2	ΔA-% μ	ΔA-% σ
Unregistered	489.26	2392.19	444.68	1153.08	2528.76	728.66	64.29	6.58	0.72	0.23	4260.86	0.00	0.00
LS 1	15.60	133.84	15.84	3.55	7.94	1.45	44.87	8.66	0.97	0.02	2150.63	5.28	8.89
LS 2	36.81	426.21	44.47	318.71	523.71	172.64	44.96	8.48	0.97	0.02	2126.81	31.75	22.22
OPT default	74.39	840.69	103.75	1207.72	2009.45	613.59	48.92	9.48	0.94	0.04	2538.84	−0.19	7.68
OPT optimal	23.89	350.99	28.67	417.90	648.24	206.70	42.83	8.65	0.97	0.02	1954.89	6.52	7.33
SIFT default	24.74	362.78	30.43	442.32	645.14	183.04	43.96	9.16	0.97	0.02	2066.20	−6.77	13.20
SIFT optimal	22.90	383.45	28.62	474.01	680.56	204.64	43.31	8.79	0.97	0.02	2001.13	−1.40	8.84
HSR	24.02	664.22	36.11	450.51	752.32	245.11	46.26	8.64	0.96	0.02	2280.25	3.18	5.32
RVSS default	93.96	4805.50	281.03	1228.69	2659.39	741.15	45.63	10.15	0.93	0.11	2072.08	−33.09	21.13
RVSS optimal	32.18	850.09	67.36	954.97	1353.44	431.53	42.46	8.89	0.96	0.04	1843.81	−8.99	5.44
ESA default	368.07	2278.21	442.01	834.71	1982.43	557.07	57.53	9.22	0.78	0.25	3127.28	0.01	0.10
ESA optimal	15.81	476.33	35.67	414.62	602.38	184.81	38.41	9.87	0.98	0.02	1603.96	2.34	2.73
MIM default	29.91	401.78	32.29	518.58	934.15	242.96	57.71	7.70	0.97	0.02	3449.70	0.01	2.38
MIM optimal	24.38	395.29	29.57	551.12	780.07	231.99	56.03	8.05	0.97	0.02	3266.80	−0.62	2.46
Voloom	39.18	730.44	48.39	713.29	1232.42	408.67	53.99	7.13	0.96	0.03	2988.03	−3.61	3.38

Prostate, high resolution

Algorithm	TRE1 μ	TRE1 max	TRE1 σ	ATRE1 μ	ATRE1 max	ATRE1 σ	RMSE μ	RMSE σ	Jaccard μ	Jaccard σ	Contrast f_2	ΔA-% μ	ΔA-% σ
Unregistered	489.25	2392.11	444.69	1152.97	2526.57	728.25	69.73	6.61	0.72	0.23	5021.08	0.00	0.00
LS 1	15.49	134.48	15.88	3.08	5.21	1.27	52.81	8.40	0.97	0.02	2939.94	4.91	8.77
LS 2	36.70	426.91	44.52	315.36	515.91	169.75	52.81	8.26	0.97	0.02	2908.40	31.28	22.08
OPT default	74.95	904.92	103.59	1327.22	2013.98	634.53	57.02	9.21	0.94	0.05	3404.82	−21.75	9.76
OPT optimal	24.25	345.68	29.46	402.79	633.01	201.36	50.75	8.43	0.97	0.02	2713.34	1.73	5.04
SIFT default	62.17	5451.71	319.97	577.46	1458.02	256.04	52.51	8.87	0.95	0.11	2838.59	−13.44	15.28
SIFT optimal	22.32	376.04	26.36	382.36	591.61	177.19	51.24	8.47	0.97	0.02	2763.28	−1.44	6.76
HSR	23.91	660.05	36.35	436.81	733.85	239.31	53.26	8.37	0.97	0.02	2990.32	1.03	5.60
RVSS default	34.35	1158.20	69.18	351.61	1070.20	148.22	50.26	9.51	0.96	0.06	2550.30	−28.06	13.25
RVSS optimal	19.49	446.90	28.31	352.14	579.83	162.65	48.92	8.56	0.97	0.02	2470.84	−4.28	3.62
ESA default	383.59	2278.27	441.44	934.43	2228.70	640.98	64.59	8.52	0.77	0.25	4043.04	0.02	0.08
ESA optimal	21.54	565.31	48.32	623.90	984.22	310.58	46.81	10.45	0.97	0.03	2346.21	1.21	2.30
MIM default	29.51	465.77	45.50	683.88	1105.29	290.42	56.74	8.12	0.96	0.03	3329.95	−0.37	3.00
MIM optimal	15.17	456.13	24.97	493.14	706.91	211.23	53.03	8.29	0.98	0.02	2944.42	−0.76	3.40
Voloom	43.35	684.11	56.28	687.46	1236.27	401.57	62.32	6.69	0.96	0.03	3945.05	−4.29	3.23

Note: Results for the unregistered images, LS based on landmarks by observer 1 (LS1) or 2 (LS2) and the automated methods (OPT, SIFT, HSR, RVSS, ESA, MIM, Voloom) using default or optimized parameters. Mean (μ), maximum (max) and standard deviation (σ) over all sections are shown. TRE and ATRE based on landmarks by observer 1 are in μm. In the online version, columns with TRE, ATRE, RMSE, *f2* and ΔA-% are colored from low (blue) to high values (red). Columns with Jaccard are colored from high (blue) to low values (red). (Color version of this table is available at *Bioinformatics* online.)

In terms of maximum TRE and ATRE, the three methods were comparable, but RVSS reached slightly lower ATRE than ESA or MIM. Among all tools, ESA and MIM also obtained the highest Jaccard index values. The RMSE and f_2 metrics do not allow comparison across different image resolutions and one should note that MIM's output was always stored at the lower resolution for technical reasons. Considering these limitations, we can observe that ESA performed best in terms of these metrics on both image resolutions ahead of RVSS. Changes in tissue area introduced by ESA, MIM and RVSS were moderate. Behind the top three, most other tools reached accuracy comparable to each other. The worst results were obtained using default parameters and for some methods, most notably ESA and RVSS, they were even comparable to the unregistered original images.

Visual examination in 3D revealed differences in the geometry of the reconstructions formed using each of the methods (Fig. 2). Compared to the undistorted reference (LS1), the distortions introduced by OPT, SIFT, HSR, ESA and MIM were a manifestation of the typical 'banana-into-cylinder' issue. This gradual straightening of curved structures is most clearly seen here in the displacement of the urethra at the top of the stacks. As indicated by the numerical ATRE values, the overall magnitude of this effect was rather similar across the tools. The distortions caused by RVSS and Voloom were more complex, representing clockwise twisting of the sample when seen from the top.

3.4 Comparison of algorithms based on the liver dataset

Results for the liver dataset are listed in Table 2. The four artificial landmarks were annotated by both observers and the two sets of TRE and ATRE values can be treated as replicates. This is reflected by linear correlation coefficients of approximately one (ranging from 0.99993 to 0.99998) for mean TRE, maximum TRE, mean ATRE and maximum ATRE calculated based on the two sets of landmarks (see Supplementary Table S1). In this case, LS thus represents an optimistic estimate of the accuracy reachable with a global affine model. Compared to the prostate sample, this dataset is more challenging to reconstruct due to the more homogeneous appearance of the tissue and the presence of deformations such as folded and torn tissue. This is reflected by the metrics, which generally indicate higher errors, except for RMSE and f_2 which are lower due to the more homogeneous image content. Ideally, it would be convenient to process different datasets without having to readjust parameters. With this in mind, we reused the parameters optimized for the prostate dataset, treating the evaluation on the liver dataset as an independent validation experiment. Based on most metrics, the optimized parameters generally resulted in an improvement over the default parameters also when applied to the liver dataset (see Table 2 and Supplementary Results).

As with the prostate, the lowest TRE values among the automated methods were achieved by ESA on the lower resolution and MIM on the high resolution data with RVSS being the third best method. The other methods reached TRE values comparable to each other. In terms of maximum TRE and ATRE, the conclusion was less clear. Voloom performed better on the lower resolution, reaching a maximum TRE second only to LS, while ESA and OPT also reached comparable values. On this dataset, MIM suffered from larger maximum errors compared to the higher quality prostate sample. The lowest mean ATRE values among all automated methods were obtained by ESA, MIM and Voloom, while in terms of maximum ATRE Voloom was superior to ESA and MIM. ESA was the top method in terms of RMSE and f_2, and MIM obtained the highest Jaccard index. Again, the poorest results were obtained when using the default values of tunable parameters.

Visualization in 3D supported the numerical results (Fig. 3). ESA, MIM and Voloom formed reconstructions with landmarks concentrated on four roughly parallel lines as expected, but some

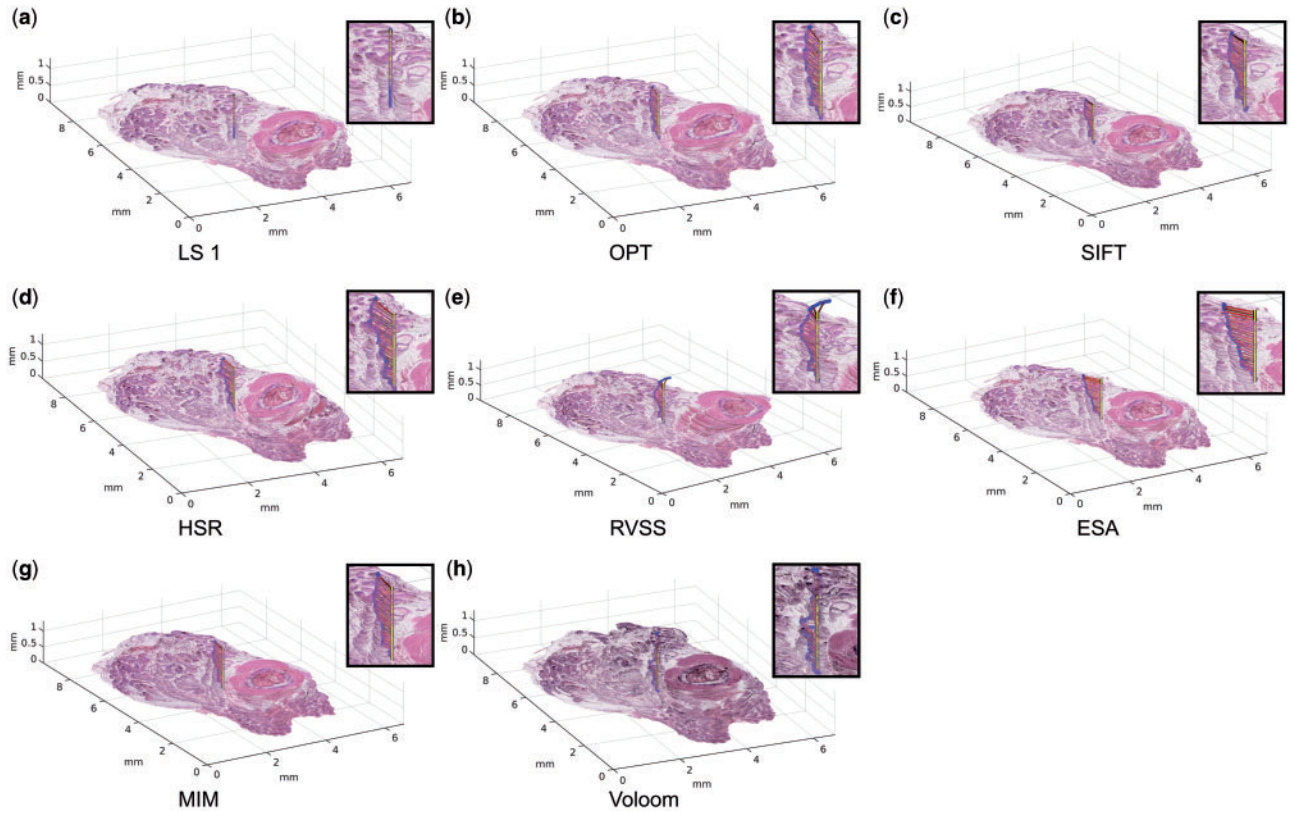

Fig. 2. Reconstructions using (**a**) LS based on landmarks by observer 1, (**b**) OPT, (**c**) SIFT, (**d**) HSR, (**e**) RVSS, (**f**) ESA, (**g**) MIM and (**h**) Voloom. Optimized parameters and the most suitable resolution were used for each method. The dots represent the trajectory of accumulated target registration error from section to section. The horizontal lines indicate the direction and magnitude of the cumulative mean displacement of each section relative to the ideal error-free trajectory (vertical line). Magnified views are shown next to each reconstruction. Viewing the high-resolution color version of the Figure online is recommended. (Color version of this figure is available at *Bioinformatics* online.)

Table 2. Evaluation results for the liver data at low (top) and high resolution (bottom)

Liver, low resolution

Algorithm	TRE1 μ	TRE1 max	TRE1 σ	ATRE1 μ	ATRE1 max	ATRE1 σ	RMSE μ	RMSE σ	Jaccard μ	Jaccard σ	Contrast f_2	ΔA-% μ	ΔA-% σ
Unregistered	726.81	2558.97	528.95	543.56	1706.62	298.02	44.90	5.03	0.67	0.15	2031.62	0.00	0.00
LS 1	27.30	396.78	55.62	25.87	314.15	35.96	34.69	6.39	0.90	0.07	1225.25	6.15	8.94
LS 2	33.52	401.27	55.70	29.52	318.41	36.55	34.76	6.41	0.90	0.07	1230.75	7.55	9.10
OPT default	200.11	1120.63	197.43	189.74	933.68	154.81	39.70	5.90	0.86	0.08	1663.83	−40.28	21.10
OPT optimal	84.86	617.62	112.51	97.28	482.65	80.44	35.26	6.44	0.92	0.06	1293.17	−10.76	8.69
SIFT default	178.38	3900.82	383.37	729.60	2096.57	511.87	36.28	7.08	0.86	0.12	1327.28	−6.61	10.43
SIFT optimal	173.15	3755.45	453.05	668.41	2837.41	572.90	35.07	6.91	0.87	0.14	1258.35	−0.78	7.44
HSR	86.99	718.85	117.16	118.15	407.31	83.00	38.27	6.26	0.92	0.06	1520.41	−15.99	9.96
RVSS default	330.02	3764.99	600.79	656.13	2186.17	494.23	36.85	7.46	0.92	0.08	1338.65	−13.23	14.70
RVSS optimal	252.32	2689.75	436.63	855.53	1677.06	334.83	35.20	7.45	0.85	0.16	1261.35	−0.39	3.31
ESA default	717.22	2558.97	539.55	538.28	1702.38	302.25	44.44	6.07	0.67	0.16	1992.03	0.00	0.01
ESA optimal	46.32	618.27	92.03	63.72	599.97	68.07	32.23	7.03	0.90	0.08	1075.18	−0.44	2.27
MIM default	121.44	2241.90	327.01	380.34	1500.07	370.61	42.83	5.70	0.90	0.11	1857.95	0.41	3.49
MIM optimal	79.74	1767.90	169.53	75.82	1233.78	108.02	42.58	5.59	0.92	0.08	1841.03	2.34	6.68
Voloom	90.98	555.46	103.81	80.12	382.78	71.12	37.69	5.39	0.91	0.07	1444.09	1.87	5.51

Liver, high resolution

Algorithm	TRE1 μ	TRE1 max	TRE1 σ	ATRE1 μ	ATRE1 max	ATRE1 σ	RMSE μ	RMSE σ	Jaccard μ	Jaccard σ	Contrast f_2	ΔA-% μ	ΔA-% σ
Unregistered	726.87	2559.07	528.92	543.55	1706.53	298.04	48.79	4.90	0.67	0.15	2396.69	0.00	0.00
LS 1	27.25	398.01	55.60	25.82	314.38	35.95	39.21	5.87	0.90	0.07	1554.89	5.87	8.92
LS 2	33.53	401.34	55.62	29.51	317.90	36.54	39.28	5.88	0.90	0.07	1560.83	7.27	9.08
OPT default	202.50	1115.20	198.27	185.80	961.31	154.84	43.85	5.48	0.86	0.08	2000.94	−40.49	20.46
OPT optimal	83.68	625.48	112.30	97.24	481.94	79.82	39.75	5.90	0.92	0.06	1628.50	−14.25	9.50
SIFT default	145.16	1388.05	173.41	223.89	1052.81	146.44	41.91	6.28	0.88	0.08	1782.81	−6.94	6.81
SIFT optimal	84.94	1026.27	130.96	157.17	630.95	117.20	39.51	6.01	0.90	0.08	1590.79	0.18	4.62
HSR	88.08	1117.63	133.55	153.43	598.88	120.99	42.24	5.73	0.92	0.07	1836.69	−19.07	10.87
RVSS default	179.82	1097.54	165.98	332.02	1052.27	165.93	42.31	5.84	0.92	0.06	1813.05	−7.96	8.40
RVSS optimal	79.26	1135.00	135.65	167.36	602.79	123.38	38.97	6.17	0.90	0.08	1548.98	−1.57	3.64
ESA default	693.75	2559.07	544.51	538.73	1711.11	301.12	47.90	6.70	0.68	0.16	2315.71	0.00	0.02
ESA optimal	60.60	929.16	142.25	56.58	832.23	99.19	37.68	6.44	0.90	0.09	1448.05	0.44	1.20
MIM default	95.74	1150.34	156.76	150.75	866.23	134.37	43.27	5.98	0.90	0.09	1896.02	0.85	3.79
MIM optimal	65.42	1060.78	122.46	66.54	646.40	78.31	42.00	5.70	0.92	0.07	1792.75	3.38	6.73
Voloom	144.08	3335.29	399.41	113.82	3159.53	274.36	42.77	4.84	0.91	0.07	1848.66	1.45	5.41

Note: Results for the unregistered images, LS based on landmarks by observer 1 (LS1) or 2 (LS2) and the automated methods (OPT, SIFT, HSR, RVSS, ESA, MIM, Voloom) using default or optimized parameters. Mean (μ), maximum (max) and standard deviation (σ) over all sections are shown. TRE and ATRE based on landmarks by observer 1 are in μm. In the online version, columns with TRE, ATRE, RMSE, *f*2 and ΔA-% are colored from low (blue) to high values (red). Columns with Jaccard are colored from high (blue) to low values (red). (Color version of this table is available at *Bioinformatics* online.)

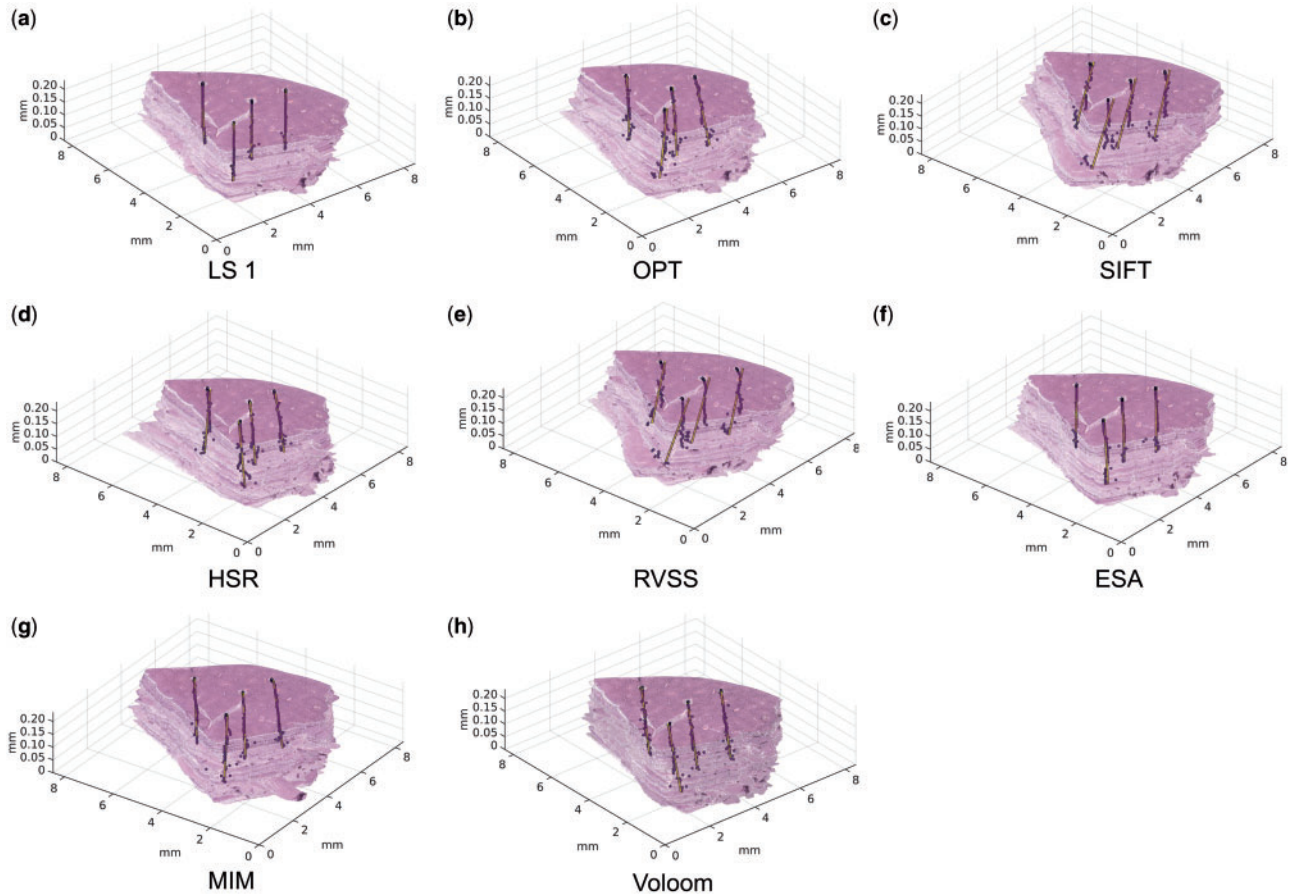

Fig. 3. Reconstructions using (**a**) LS based on landmarks by observer 1, (**b**) OPT, (**c**) SIFT, (**d**) HSR, (**e**) RVSS, (**f**) ESA, (**g**) MIM and (**h**) Voloom. Optimized parameters and the most suitable resolution were used for each method. The locations of the four landmark points on each section are indicated with dots, shown together with lines of best fit to each of the four series of points. Note that the scale of the vertical axis is different from the horizontal axes in the visualization. Viewing the high-resolution color version of the Figure online is recommended. (Color version of this figure is available at *Bioinformatics* online.)

distortion is visible at the bottom part of the stack reconstructed by MIM. These kind of distortions were more severe in the case of OPT, SIFT, HSR and RVSS.

4 Discussion

Based on this study, methods utilizing locally varying transformations (ESA, MIM, RVSS, Voloom) were superior to those constrained to global affine models (OPT, SIFT, HSR). ESA was the only method to consistently outperform or match the other approaches on two datasets based on the majority of metrics. In the case of the higher quality prostate dataset, differences in accuracy between the tools were rather subtle. All three top-performing methods on this dataset incorporate an elastic transformation model: MIM and RVSS use a B-spline grid and ESA is based on a piecewise linear mesh. While methods relying on a global transformation model also performed reasonably well, the additional accuracy offered by elastic transformations could be crucial when microstructure at the cellular scale is of interest. In the case of the liver sample, more profound differences between the methods were observed, likely due to the more challenging tissue content and the presence of deformations, which cannot be compensated for using a global model. ESA, MIM and Voloom stood out from the other methods. While Voloom appeared to be less accurate on average compared to ESA and MIM based on mean TRE, it demonstrated the lowest maximum and accumulated errors of all automated methods, indicating capability to avoid propagation of errors even in the presence of considerable deformations. The ability of the algorithms to tolerate such deformations is a significant benefit. Due to the mostly manual nature of histological sectioning and brittleness of the thin tissue sections, deformations in the form of folds and tears often occur. This challenge is especially encountered in 3D histology, when uninterrupted sequences of sections are desired.

Another important property of algorithms to consider is sensitivity to adjustable parameters. Even an algorithm that produces highly accurate results with a carefully selected set of parameter values will be useless if the user has little chance of finding this set of values. Comparing algorithms from this perspective is difficult. Each algorithm has a different set of parameters and the range of values to evaluate has to be selected for each parameter, which can in turn affect the amount of variation observed in the results. Nevertheless, this study still provides a semi-quantitative view of the sensitivity of the studied algorithms against parameter adjustments. Of the evaluated methods, LS, HSR and Voloom are the most convenient due to their lack of tunable parameters. OPT and SIFT also produced similar results with most parameter values. The results produced by ESA varied greatly depending on parameters, but we discovered numerous combinations leading to almost optimal results. In the case of MIM, there are only a handful of tunable parameters and they are relatively easy to tune. Moreover, ESA and MIM appear to

be well-behaving in the sense that parameters optimized for the prostate dataset also suited the liver dataset. In contrast, RVSS was found to be difficult to optimize and even though its accuracy using optimized settings was close to ESA and MIM on the prostate dataset, reaching this level of accuracy without automated parameter tuning would be challenging.

An open question common to all of the methods is how image resolution affects reconstruction accuracy. A pixel size close to the section spacing is often recommended (Amunts *et al.*, 2013; Braumann *et al.*, 2005; Dauguet *et al.*, 2007; Ju *et al.*, 2006; Kartasalo *et al.*, 2016; Saalfeld *et al.*, 2012) based on the assumption that objects smaller than this are only visible on a single section and are thus not useful for registration, and may even introduce errors (Beare *et al.*, 2008). However, suitably oriented elongated structures such as blood vessels can be observed on several sections even if their diameter on the image plane is smaller than the section spacing. In principle, some algorithms might thus benefit from a smaller pixel size. We evaluated reconstruction accuracy using pixel sizes of 3.68 and 7.36 μm. Based on the rule of thumb above, it is unclear which one of these should be preferred given a section spacing of 5 μm. Our results indicate that using a pixel size close to the section spacing is a reasonable starting point, but the optimal image resolution depends on the algorithm and also somewhat on the image content. Furthermore, we cannot rule out the possibility that algorithms which performed better on the high resolution images, most notably MIM, might benefit from an even smaller pixel size. In conclusion, the image resolution thus needs to be selected experimentally for each application and algorithm.

The two samples selected for this study are markedly different in their histological composition. The fact that the top methods performed well on both the prostate and the liver dataset without any retuning of parameters indicates that these methods are not overly sensitive to tissue appearance, and that the results obtained in this study are not specific to a single dataset. However, some variation in the relative performance of the algorithms on the two datasets was still observed. Thus, collecting and annotating additional datasets representing diverse tissue types and other histological stainings, such as immunohistochemistry, remains an important goal for future studies.

While we evaluated a comprehensive set of methods for 3D histology, it might be worthwhile to adapt general-purpose image registration algorithms to this context. Another opportunity, not supported by any of the methods here, could be the exploitation of additional data obtained e.g. by magnetic resonance imaging or in the form of blockface images (Amunts *et al.*, 2013; Casero *et al.*, 2017; Dauguet *et al.*, 2007; Gibson *et al.*, 2013; Johnson *et al.*, 2010; Stille *et al.*, 2013). Furthermore, although advances in image acquisition and processing have enabled the first steps towards 3D histology, sample preparation still constitutes a significant bottleneck. In the future, emerging technologies for automated sample preparation (Onozato *et al.*, 2011) or integrated sectioning and imaging (Li *et al.*, 2010; Ragan *et al.*, 2012) might potentially transform 3D histology into a high-throughput process.

Acknowledgements

We thank Ignacio Arganda-Carreras, Martin Groher, Derek Magee, Stephan Saalfeld and Ved Sharma for their helpful advice. Katja Liljeström, Marja Pirinen and Marika Vähä-Jaakkola are acknowledged for skillful technical assistance.

Funding

This work was supported by Academy of Finland [269474]; Tekes [269/31/2015]; Cancer Society of Finland; Emil Aaltonen Foundation; Finnish Foundation for Technology Promotion; KAUTE Foundation; and Orion Research Foundation.

Conflict of Interest: none declared.

References

Amunts,K. *et al.* (2013) BigBrain: an ultrahigh-resolution 3D human brain model. *Science*, **340**, 1472–1475.

Arganda-Carreras,I. *et al.* (2006) Consistent and elastic registration of histological sections using vector-spline regularization. In: *International Workshop on Computer Vision Approaches to Medical Image Analysis*, pp. 85–95.

Arganda-Carreras,I. *et al.* (2010) 3D reconstruction of histological sections: application to mammary gland tissue. *Microsci. Res. Technol.*, **73**, 1019–1029.

Beare,R. *et al.* (2008) An assessment of methods for aligning two-dimensional microscope sections to create image volumes. *J. Neurosci. Methods*, **170**, 332–344.

Braumann,U. *et al.* (2005) Three-dimensional reconstruction and quantification of cervical carcinoma invasion fronts from histological serial sections. *IEEE Trans. Med. Imaging*, **24**, 1286–1307.

Bull,A.D. (2011) Convergence rates of efficient global optimization algorithms. *J. Mach. Learn. Res.*, **12**, 2879–2904.

Calì,C. *et al.* (2016) Three-dimensional immersive virtual reality for studying cellular compartments in 3D models from EM preparations of neural tissues. *J. Comp. Neurol.*, **524**, 23–38.

Cardona,A. *et al.* (2012) TrakEM2 software for neural circuit reconstruction. *PloS One*, **7**, e38011.

Casero,R. *et al.* (2017) Transformation diffusion reconstruction of three-dimensional histology volumes from two-dimensional image stacks. *Med. Image Anal.*, **38**, 184–204.

Cifor,A. *et al.* (2011) Smoothness-guided 3-D reconstruction of 2-D histological images. *Neuroimage*, **56**, 197–211.

Dauguet,J. *et al.* (2007) Three-dimensional reconstruction of stained histological slices and 3D non-linear registration with in-vivo MRI for whole baboon brain. *J. Neurosci. Methods*, **164**, 191–204.

Fischler,M.A. and Bolles,R.C. (1981) Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, **24**, 381–395.

Fitzpatrick,J.M. *et al.* (1998) Predicting error in rigid-body point-based registration. *IEEE Trans. Med. Imaging*, **17**, 694–702.

Gaffling,S. *et al.* (2015) A Gauss-Seidel iteration scheme for reference-free 3-D histological image reconstruction. *IEEE Trans. Med. Imaging*, **34**, 514–530.

Ghaznavi,F. *et al.* (2013) Digital imaging in pathology: whole-slide imaging and beyond. *Annu. Rev. Pathol.-Mech.*, **8**, 331–359.

Gibson,E. *et al.* (2013) 3D prostate histology image reconstruction: quantifying the impact of tissue deformation and histology section location. *J. Path. Inform.*, **4**, 31.

Griffin,J. and Treanor,D. (2017) Digital pathology in clinical use: where are we now and what is holding us back? *Histopathology*, **70**, 134–145.

Haralick,R.M. and Shanmugam,K. (1973) Textural features for image classification. *IEEE Trans. Syst. Man Cybern.*, **3**, 610–621.

Hiner,M.C. *et al.* (2016) ImageJ-MATLAB: a bidirectional framework for scientific image analysis interoperability. *Bioinformatics*, **33**, 629–630.

Hughes,C. *et al.* (2013) Robust alignment of prostate histology slices with quantified accuracy. *IEEE Trans. Biomed. Eng.*, **60**, 281–291.

Johnson,G.A. *et al.* (2010) Waxholm space: an image-based reference for coordinating mouse brain research. *Neuroimage*, **53**, 365–372.

Jones,D.R. (2001) A taxonomy of global optimization methods based on response surfaces. *J. Global. Optim.*, **21**, 345–383.

Ju,T. *et al.* (2006) 3D volume reconstruction of a mouse brain from histological sections using warp filtering. *J. Neurosci. Methods*, **156**, 84–100.

Kartasalo,K. *et al*. (2016) Benchmarking of algorithms for 3D tissue reconstruction. In: *2016 IEEE International Conference on Image Processing (ICIP)*, pp. 2360–2364.

Koos,B. *et al*. (2015) Next-generation pathology—surveillance of tumor microecology. *J. Mol. Biol*., **427**, 2013–2022.

Ledford,H. (2017) Cell atlases race to map the body. *Nature*, **542**, 404–405.

Lein,E.S. *et al*. (2007) Genome-wide atlas of gene expression in the adult mouse brain. *Nature*, **445**, 168–176.

Li,A. *et al*. (2010) Micro-optical sectioning tomography to obtain a high-resolution atlas of the mouse brain. *Science*, **330**, 1404–1408.

Lowe,D.G. (2004) Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis*., **60**, 91–110.

Magee,D. *et al*. (2015) Histopathology in 3D: from three-dimensional reconstruction to multi-stain and multi-modal analysis. *J. Path. Inform*., **6**, 6.

Malandain,G. *et al*. (2004) Fusion of autoradiographs with an MR volume using 2-D and 3-D linear transformations. *Neuroimage*, **23**, 111–127.

Meijering,E. *et al*. (2016) Imagining the future of bioimage analysis. *Nat. Biotechnol*., **34**, 1250–1255.

Mignardi,M. *et al*. (2017) Bridging histology and bioinformatics—computational analysis of spatially resolved transcriptomics. *Proc. IEEE*, **105**, 530–541.

Onozato,M.L. *et al*. (2011) Evaluation of a completely automated tissue-sectioning machine for paraffin blocks. *J. Clin. Pathol*., 200205.

Onozato,M.L. *et al*. (2012) A role of three-dimensional (3D)-reconstruction in the classification of lung adenocarcinoma. *Anal. Cell. Pathol*., **35**, 79–84.

Paish,E.C. *et al*. (2009) Three-dimensional reconstruction of sentinel lymph nodes with metastatic breast cancer indicates three distinct patterns of tumour growth. *J. Clin. Pathol*., **62**, 617–623.

Ragan,T. *et al*. (2012) Serial two-photon tomography for automated ex vivo mouse brain imaging. *Nat. Methods*, **9**, 255–258.

Roberts,N. *et al*. (2012) Toward routine use of 3D histopathology as a research tool. *Am. J. Pathol*., **180**, 1835–1842.

Rohlfing,T. (2012) Image similarity and tissue overlaps as surrogates for image registration accuracy: widely used but unreliable. *IEEE Trans. Med. Imaging*, **31**, 153–163.

Rojas,K.D. *et al*. (2015) Methodology to study the three-dimensional spatial distribution of prostate cancer and their dependence on clinical parameters. *J. Med. Imaging*, **2**, 037502.

Rusk,N. (2016) Genomics: spatial transcriptomics. *Nat. Methods*, **13**, 710–711.

Saalfeld,S. *et al*. (2012) Elastic volume reconstruction from series of ultra-thin microscopy sections. *Nat. Methods*, **9**, 717–720.

Schindelin,J. *et al*. (2012) Fiji: an open-source platform for biological-image analysis. *Nat. Methods*, **9**, 676–682.

Schneider,C.A. *et al*. (2012) NIH Image to ImageJ: 25 years of image analysis. *Nat. Methods*, **9**, 671.

Shahriari,B. *et al*. (2016) Taking the human out of the loop: a review of bayesian optimization. *Proc. IEEE*, **104**, 148–175.

Snoek,K. *et al*. (2012) Practical Bayesian optimization of machine learning algorithms. *Adv. Neurol. Int*., 2951–2959.

Song,Y. *et al*. (2013) 3D reconstruction of multiple stained histology images. *J. Path. Inform*., **4**, 7.

Sotiras,A. *et al*. (2013) Deformable medical image registration: a survey. *IEEE Trans. Med. Imaging*, **32**, 1153–1190.

Ståhl,P.L. *et al*. (2016) Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science*, **353**, 78–82.

Stille,M. *et al*. (2013) 3D reconstruction of 2D fluorescence histology images and registration with in vivo MR images: application in a rodent stroke model. *J. Neurosci. Methods*, **219**, 27–40.

Studholme,C. *et al*. (1999) An overlap invariant entropy measure of 3D medical image alignment. *Pattern Recognit*., **32**, 71–86.

Teodoro,G. *et al*. (2017) Algorithm sensitivity analysis and parameter tuning for tissue image segmentation pipelines. *Bioinformatics*, **33**, 1064–1072.

Theart,R.P. *et al*. (2017) Virtual reality assisted microscopy data visualization and colocalization analysis. *BMC Bioinformatics*, **18**, 64.

Thevenaz,P. *et al*. (1998) A pyramid approach to subpixel registration based on intensity. *IEEE Trans. Image Process*, **7**, 27–41.

Wang,Y. *et al*. (2015) Three-dimensional reconstruction of light microscopy image sections: present and future. *Front. Med*., **9**, 30–45.

Xu,Y. *et al*. (2015) A method for 3D histopathology reconstruction supporting mouse microvasculature analysis. *PloS One*, **10**, e0126817.