Markerless 3D spatio-temporal reconstruction of microscopic swimmers from video

Felix Salfelder, Omer Yuval, Thomas P. Ilett, David C. Hogg, Thomas Ranner, Netta Cohen
University of Leeds
School of Computing
Leeds, United Kingdom
{scsfsa,scoy,scti,D.C.Hogg,T.Ranner,N.Cohen}@leeds.ac.uk

***Abstract*—3D object reconstruction of deformable objects is a long standing challenge for computer vision. Here we develop a system for the 3D reconstruction of a single marker-less object – a freely moving biological swimmer in 3D space – using a passive, fixed-camera set-up. We focus on microscopic, long and thin (1 mm long, 80 µm thick) roundworms. Our set-up provides the resolution required both to track the animal's coordinates across a large volume and to reconstruct its 3D posture at every frame. A data-pipeline is presented which combines model calibration, 2D image analysis and 3D reconstruction of the body midline, representing the complete posture up to orientation and internal twist. We present results, validation and open challenges, including instances of occlusion due to insufficient projected information, and experimental limitations of resolution and focus.**

I. Background

The study of animal behavior spans a number of related challenges. Trajectories of animals in space help to inform models of decision making, strategies for predation, spatial mapping and navigation. Underpinning these trajectories are motor behaviors, principally locomotion which, in swimmers, arises from internal control of postures and kinematics, subject to the body biomechanics and fluid dynamics. Capturing the range of spatial information within a single data set requires an imaging setup that can be used to resolve postures within a sufficiently large field of view so as to capture long trajectories. When the behavior is fundamentally three dimensional and not well approximated by movement on a surface, fixed-camera, passive imaging systems must balance between the resolution and depth of field. High resolution imaging tends therefore to focus on very short durations. Here, we describe a system capable of imaging the locomotion of the nematode (roundworm) *Caenorhabditis elegans* over many minutes, and propose a computer vision system capable of reconstructing both 3D body shapes and trajectories of a freely moving animal in 3D volumes.

C. elegans locomotion has been the focus of a large body of work, almost all of which is restricted to planar motion, although interest in 3D locomotion has recently increased [1]. The animal moves primarily by propagating sinusoidal undulations along the body, opposite to the direction of motion (either forward or backward). On a surface, it steers by biasing the undulations on either side of the body, and it performs occasional maneuvers to undertake sharp turns, the most common of which is dubbed an 'omega turn' [2]. In contrast, the reconstruction of postures and characterization of behaviors in 3D are, as yet, open questions [3]. Due to the size of the worm ($\sim 1\,\mathrm{mm}$ in length), optical effects such as distortion and relative positioning are no longer negligible, and protocols must therefore address dynamic calibration problems.

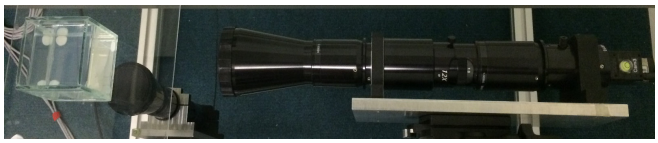

Fig. 1. One axis of the imaging setup including a telecentric lens pointing at a fluid-filled cube.

Here, we focus on microscopic worms: freely bending and twisting, long and slender objects. For our purposes the length and local diameter of the swimmer is assumed to be fixed in time and hence of fixed volume. Despite the transparency of *C. elegans*, it can be (partially) self-occluding, and, because there is a resource limit to the number of simultaneous projections which can be collected, there can be (rare) instances in which all projections are far from ideal.

Multiple tools have been developed to reconstruct the posture of *C. elegans* in 2D [4]–[8] (see [9] for a review). Most of these methods are based on a combination of 'shrinking and pruning' algorithms which were first formulated in the prairie-fire model [10]. The images are transformed into a silhouette of the worm which can then be shrunk to a skeleton using standard methods. 3D spatio-temporal reconstruction from multiple views similarly often relies on silhouettes of the 2D projection images. In principle for convex objects, intersected back-projections of a sufficient number of silhouettes obtained from multiple projections can recover a volumetric description.

In our setup, wild-type *C. elegans* worms are placed in a 3D glass cube filled with transparent gelatinous fluids of various concentrations. Our current study is limited to the volumetric reconstruction of wild-type worms, in which self-occlusion, e. g. due to coiling body shapes, becomes the dominant challenge for both machine and human vision. Our imaging setup consists of 3 fixed, passive, nearly orthogonal views of our sample using telecentric lenses (Fig. 1), with back-lighting in each direction (as the worms are transparent). The imaging setup is illustrated in Fig. 2. The quality of

the microscopic imagery is limited by the trade-offs between the magnification, numerical aperture and depth of field. For biological reasons, only red light is suitable, and plays against the achievable optical resolution due to its high wavelength. With worms undulating at up to 2 Hz, and a sampling at 25-40 frames per second, a further constraint on the quality of reconstruction is given by the difference in posture between consecutive frames.

We tested two additional approaches: (i) 3D videography, using 3 microscopic cameras focused on the swimmer, mounted on a motorized frame with 3 orthogonal degrees of freedom controlled with real time closed loop tracking, but hardware lacked robustness for extended recordings (losing the object if tracking along any of the axes momentarily failed) (ii) 3D holographic imaging [11]. We found the holographic approach lacked sufficient resolution given the low magnification required to capture the entire field of view.

II. Outline

Our reconstruction approach relies on photogrammetry: a set of methods for reconstructing three dimensional objects from planar camera images. For rigid objects, photogrammetry typically relies on triangulation of a set of control points. Generalizing such approaches for spatio-temporal reconstruction of deformable objects typically requires tracking a set of features or markers on the body. In our case, as *C. elegans* worms are radially symmetric along most of the body, the magnification is necessarily too limited to resolve local anatomical features such as the vulva. We are also unable to identify features of the animal's internal or reference coordinate frame other than the tip of head and tail based on the fuzzy shadow-like images. Capturing only calibration images and triplets of 2D grayscale projection images, our reconstruction aim is therefore restricted to determining the midline of the body, represented as a curve in 3D, over time.

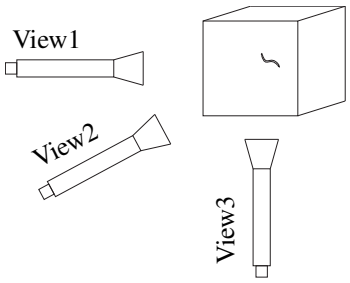

Fig. 2. Experimental setup schematic.

An example of a grayscale camera image is shown in Fig. 3. The goal of the reconstruction is a curve along the centerline of the animal in real-world coordinates. The reconstruction procedure is as follows:

a) Calibration of the camera setup using calibration images taken before the experiment.
b) Image normalization, object tracking and triangulation.
c) 2D image segmentation to find midlines using a trained equivariant convolutional neural network.

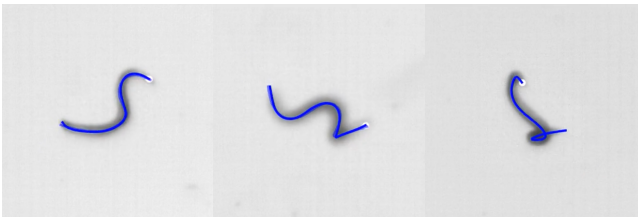

Fig. 3. Close-up of raw camera views with projected reconstruction, 200×200 pixels cropped from 2048 × 2048 8bpp microscope image.

d) Correlation-based fine tuning of the camera calibration along the moving object.
e) Space carving with three-way majority voting to obtain a discrete skeleton.
f) Curve fit optimization using a finite element formulation and weighted candidate points.

III. Methods

a) Calibration: To map a set of projection pixels to the three dimensional metric space, we first parametrize our camera model. The calibration protocol is performed on images of an oriented grid of circular feature points in many (>100) different configurations captured simultaneously in all three camera views (Fig. 4). Due to technical reasons, these images must be taken hours before the video recording. First, intrinsic parameters (the focal length and lens distortion), are established for each camera individually. Each view is modeled as a single pinhole camera A bundle adjustment algorithm is used to fit extrinsic parameters that describe where each camera is located relative to the sample. The pattern detection and calibration routines are performed using customized C++ code derived from the OpenCV software library [12]. Using this algorithm we typically obtain root mean squared (RMS) reprojection error on the feature points of less than 10 pixels. Both the intrinsic and extrinsic camera model parameters may change in the hours between the initial calibration and the experiment, e. g. due to the setting of our gel-based fluid. To account for these changes, we recalibrate the extrinsic parameters further as described below (d).

b) Normalization: In the recordings of the swimming animals, the transparent worm appears dark on a light gray background. In most cases, the background is virtually static over time, with an RMS noise level of around $\pm 1\%$. Exceptions occur due to transient appearances of bubbles in the medium. For each camera we obtain a single background image per clip (typically minutes long) from the maximum values of temporally low-pass filtered pixel intensities. The background image is subtracted from the video frames and the residual intensities are then normalized in each frame.

c) Segmentation: To generate informative 2D projections the frames are passed through a convolutional neural network (CNN) [13] to produce a triplet of initial 2D midline estimates before this information is lifted and combined in 3D. We have implemented a CNN with 28 layers that processes 200×200

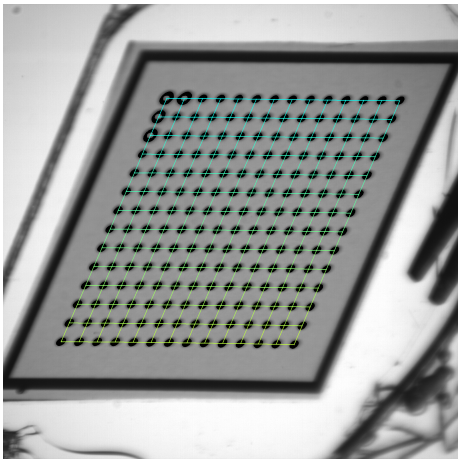

Fig. 4. Calibration image (2048×2048) showing grid of about $1\,\mathrm{cm}^2$ in size. Green overlay: detected structure used in the calibration algorithm.

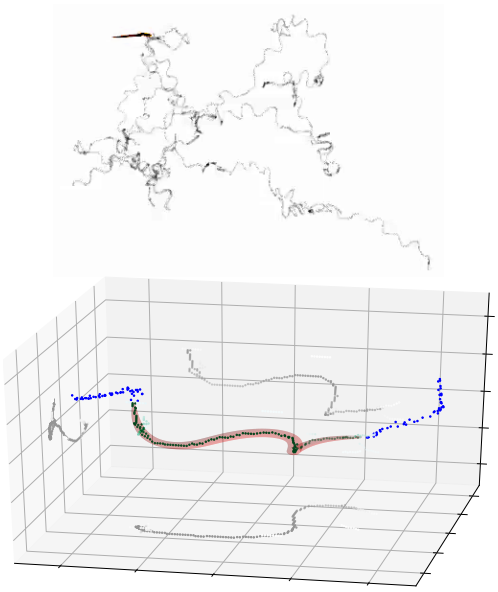

Fig. 5. Top: 13 minute trajectory to scale with worm size (about $1\mathrm{mm}$). Bottom: Shape reconstruction showing weighted control points (green) with projections. The blue dots show the respective past and future head and tail trajectories to indicate reconstruction jitter.

floating point 2D-arrays and outputs arrays of the same dimension. The first half of the network gradually downsamples the data using strided convolutions, the other half uses the corresponding 'inverse' convolution layers for upsampling. Three convolution layers are usually followed by a linear rectification unit (ReLU). The CNN was trained on a set of 120 unrelated 2D images for which the midlines have been hand-annotated. This curated data set was constructed so as to cover the range of observed postures with and without occlusions. As the training loss function we use the ℓ_2 distance between hand annotations and output, after masking a two pixel grace area around the annotation. The masking compensates for spatial inaccuracies in the manual annotations and avoids over-fitting. Our filters are equivariant with respect to mirroring and 90° rotation. Equivariant networks are constructed such that similar inputs produce similar outputs, and can be trained without the need for data augmentation. They have been proposed and used in classifiers [14].

With this network, we obtain triplets of gray-scale images each highlighting a proposed midline of the animal in its respective plane. To sift midline pixels, we post-process the output of the CNN. This step consists of applying a high-pass filter with a cut-off on the set of candidate midline points followed by selection of the largest connected component. The inferred midlines are about 2 pixels wide, but often include imperfections such as branching or loss due to occlusion or limited contrast. These imperfections most frequently occur near the head and tail ends where the (mostly transparent) worm tapers out. Our approach resolves these uncertainties when the three views are combined.

d) Local re-calibration: To improve the geometry of the camera setup near the object of interest, we consider the visual overlap of the images seen in the three nearly orthogonal camera views. Starting from the initial global calibration we use a model of light rays crossing the object in a Cartesian cube centered around the object of interest. Using stochastic gradient descent, we optimize the three-way correlation of the brightness with respect to shifts along the local coordinate axes. This optimization is performed on batches of 20 frame triplets, at the original resolution, and repeated every 10 frames. Our implementation with pytorch [15] makes use of gradient back-propagation and extensive parallel processing across GPUs.

After temporal smoothing we obtain a camera model with 1-2 pixel accuracy.

e) Skeletonizing: Given a triplet of recalibrated binary images showing midlines of a worm from three views, we now choose the 3D voxels that best match the midlines in all three views. We define this selection of voxels in the Cartesian cube as the discrete skeleton, centered around the object constructed for calibration.

We compute thin skeletons of the 2D midline segments using the Guo-Hall algorithm [16]. These 2D skeletons give rise to a thin 3D skeleton constructed as the Lee skeleton [17] of the voxels common in the 3D lifts of the 2D skeletons. A thin 3D skeleton is a 3D shape approximating the desired set of midline voxels from below, but it lacks voxels that are missing from one or more of the projections. This loss is due to inaccuracies in calibration and in the 2D image segmentation, especially towards the transparent head and tail regions. To compensate for the loss, for each voxel, we consider the three pixels in the projections it maps to. We recover voxels that map into at least two of the regions close to the midlines, so long as these include the thin 2D skeleton at least once.

f) Curve fitting: For each time point, we fit an arc-length parametrized 3D curve to the skeleton with control points at the centers of the voxels obtained in the previous step. The curve is modeled as a 1D mesh representing an elastic rod in a fluid with an internal stiffness using a finite

element formulation [18]. Control points are weighted by the product of the corresponding respective pixel intensities. Forces determine the position and shape of the curve. Each control point pulls on the nearest mesh point with a spring force proportional to the distance and the weight of the control point. Similarly each mesh point is pulled to its nearest control point. In a simulation, the model rod immersed in a fluid grows over time to a predetermined full length while giving in to the forces. In the time limit, we obtain a reconstructed midline for a single frame as a local minimum of the energy in the stiffness and applied forces. A shortened curve serves as the initial position of the rod the subsequent frame. In this way, the orientation of the curve is propagated through time.

g) Applying the pipeline: We have successfully processed about one hour of video recordings. These include a range of different magnifications and animals swimming in a range of different fluids. An example trajectory along with posture reconstruction is shown in Fig. 5.

IV. Summary

The study of microswimmer locomotion in 3D presents many technical and experimental challenges. Here we present a detailed approach for recovering not just the trajectory of the swimmer in a large volume but also parametrized 3D postures at every point in time. Our method takes synchronized microscopic videos together with calibration images. It includes a camera model with a two-phase calibration procedure, a convolutional neural network trained with a hand-curated data set to identify the worm in each 2D image, and finally a numerical optimization scheme to estimate the shape. The feature detection is parallelized through batch processing, and the recalibration optimization uses multiple GPUs in parallel.

Many of the challenges addressed in this work are well known to researchers interested in studying 3D motility and behavior of swimmers or flyers. Indeed, in addition to often costly equipment, the absence of well established and reliable image analysis pipelines makes it hard to obtain and prepare data for further analysis. This presents a significant barrier to entry for new researchers and is undoubtedly one of the main reasons why many species are overwhelmingly studied only in 2D. The 2D space however, while more accessible for the experimentalist, frequently does not capture the natural environment or behavior of the subject.

Here we have presented and demonstrated a pipeline that we have validated for about 1 hour of video footage. Analysis of this data will fuel a variety new biological and biophysical insights on animal behavior, biomechanics and active swimming more generally. This paper presents the methods, and the data will be made available upon publication of the analysis results. While relatively robust, a limitation of the pipeline includes a slow recovery after poorly-reconstructed or failed frames. Further work is under way to improve the robustness, speed and automation of this pipeline. It is our hope that this proposed methodology will provide a reliable outline for similar studies and will assist in supporting the research on small species in 3D environments.

Acknowledgements

This research was supported by EPSRC (EP/J004057/1 and EP/S01540X/1) and by a Leverhulme Trust Early Career Fellowship (TR). The work was undertaken on ARC4, part of the HPC facilities at the University of Leeds, UK and on JADE, a UK Tier-2 resource, funded by EPSRC, owned by the University of Oxford and hosted at the Hartree Centre.

References

[1] A. Bilbao, A. K. Patel, M. Rahman *et al.*, "Roll maneuvers are essential for active reorientation of *C. elegans* in 3d media," *Proceedings of the National Academy of Sciences*, vol. 115, no. 16, pp. E3616–E3625, 2018.

[2] J. M. Gray, J. J. Hill, and C. I. Bargmann, "A circuit for navigation in caenorhabditis elegans," *Proceedings of the National Academy of Sciences*, vol. 102, no. 9, pp. 3184–3191, feb 2005. [Online]. Available: https://doi.org/10.1073%2Fpnas.0409009101

[3] M. Shaw, H. Zhan, M. Elmi *et al.*, "Three-dimensional behavioural phenotyping of freely moving *C. elegans* using quantitative light field microscopy," *PLOS ONE*, vol. 13, no. 7, 2018.

[4] Z. Feng, C. J. Cronin, J. H. Wittig, P. W. Sternberg, and W. R. Schafer, "An imaging system for standardized quantitative analysis of *c. elegans* behavior," *BMC Bioinformatics*, vol. 5, p. 115, 2004.

[5] W. Geng, P. Cosman, C. C. Berry, Z. Feng, and W. R. Schafer, "Automatic tracking, feature extraction and classification of *c. elegans* phenotypes," *IEEE Trans. Biomed. Eng.*, vol. 51, pp. 1811–1820, 2004.

[6] G. Stephens, B. Johnson-Kerner, W. Bialek, and W. Ryu, "Dimensionality and dynamics in the behavior of *c. elegans*," *PLoS Comput. Biol.*, vol. 4, p. e1000028, 2008.

[7] S. Berri, J. H. Boyle, M. Tassieri, I. A. Hope, and N. Cohen, "Forward locomotion of the nematode *C. elegans* is achieved through modulation of a single gait," *HFSP Journal*, vol. 3, no. 3, pp. 186–193, 2009, pMID: 19639043. [Online]. Available: https://doi.org/10.2976/1.3082260

[8] S. Nagy, M. Goessling, Y. Amit, and D. Biron, "A generative statistical algorithm for automatic detection of complex postures," *PLoS Comput Biol*, vol. 11, no. 10, p. e1004517, 2015.

[9] S. J. Husson, W. S. Costsa, C. Schmitt, and A. Gottshalk, "Keeping tack of worm trackers," in *WormBook*, T. *C. elegans* Research Community, Ed., 2012. [Online]. Available: http://www.wormbook.org.

[10] H. Blum, "A transformation for extracting new descriptors of shape," in *Models for the Perception of Speech and Visual Form*, W. Wathen-Dunn, Ed. MIT Press, 1967, pp. 362–380.

[11] F. Cheong, S. Duarte, S. Lee *et al.*, "Holographic microrheology of polysaccharides from streptococcus mutans biofilms," *Rheol Acta 48*, pp. 109–115, 2009.

[12] "Camera calibration With OpenCV," https://docs.opencv.org/3.1.0/d4/d94/tutorial_camera_calibration.html, accessed: 2018-01-08.

[13] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, http://www.deeplearningbook.org.

[14] T. Cohen and M. Welling, "Group equivariant convolutional networks," in *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, ser. JMLR Workshop and Conference Proceedings, M. Balcan and K. Q. Weinberger, Eds., vol. 48. JMLR.org, 2016, pp. 2990–2999.

[15] A. Paszke, S. Gross, F. Massa *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle *et al.*, Eds. Curran Associates, Inc., 2019, pp. 8024–8035. [Online]. Available: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

[16] Z. V. Guo and L. Mahadevan, "Limbless undulatory propulsion on land," *Proceedings of the National Academy of Sciences*, vol. 105, no. 9, pp. 3179–3184, feb 2008.

[17] T. Lee, R. Kashyap, and C. Chu, "Building skeleton models via 3-d medial surface axis thinning algorithms," *CVGIP: Graphical Models and Img Processing*, vol. 56, no. 6, pp. 462 – 478, 1994. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S104996528471042X

[18] T. Ranner, "A stable finite element method for low inertia undulatory locomotion in three dimensions," *Applied Numerical Mathematics*, vol. 156, pp. 422 – 445, 2020. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0168927420301537