HusMorph: A simple machine learning app for automated morphometric landmarking

Henning H. Kristiansen[1,2,*], Moa Metz[1], Lorena Silva-Garay[1], Fredrik Jutfelt[1,3,#], Robine H.J. Leeuwis[1,#]

[1] Department of Biology, Norwegian University of Science and Technology, Høgskoleringen 5, 7034, Trondheim, Norway.
[2] Department of Computer Technology and Informatics, Norwegian University of Science and Technology, Høgskoleringen 5, 7034, Trondheim, Norway.
[3] Department of Biological and Environmental Sciences, University of Gothenburg, Medicinaregatan 7B, 41390, Göteborg, Sweden.

*Corresponding author: Department of Biology, Department of Computer Technology and Informatics. Email: skihenning@gmail.com.
#F.J. and R.H.J.L. are co-senior authors and have contributed equally to this work.

Email addresses: lorena.silvagaray@gmail.com (L.S.G.), rhjleeuwis@gmail.com (R.H.J.L.), fredrik.jutfelt@bioenv.gu.se (F.J.)

Author’s ORCID numbers:
M.M 0009-0002-4397-3948
L.S.G. 0000-0002-9332-6311
R.H.J.L.0000-0002-6687-4304
F.J. 0000-0001-9838-3991

Abstract

Manually obtaining the length and other morphometric features of an animal can be time-consuming, and consistent measurements are challenging with large datasets. By leveraging high-throughput computing power and machine learning-based computer vision, such phenotypic data can be rapidly collected with high accuracy. Here we present HusMorph, a novel application with a simple and intuitive graphical user interface (GUI), based on the same machine learning method used in other pipelines such as ML-morph. It consists of an all-in-one package with the goal of making machine learning easy to use for non-experts. The user starts by setting any number of landmarks on a set of photos captured with a standardized setup. From this set, a machine learning model is generated by automatically and randomly searching for the best performing parameters. Next, the user can apply the model to predict landmarks on new standardized photos, and visually confirm and export the results of the predictions. For measuring length between landmarks, an additional feature allows for detecting a scale bar for each photo to convert the length from pixels to a metric unit. Our application has been validated and applied to extract standard length from 1,935 photos of zebrafish and performs with about 99.5% accuracy compared to manual measurements along with 100% scale bar detection.

Keywords: artificial intelligence (AI), automation, images, morphometrics, phenotyping, user-friendly

Introduction

Morphometric measurements are essential in most disciplines of biology, and important for advancing our understanding of evolution, species identification and taxonomy, genotype-phenotype relationships, effects of various factors and environmental stressors on body condition, biomechanics, and paleontology (Klingenberg, 2010; Rohlf & Marcus, 1993; Rowiński et al., 2015). In recent years, there is a growing trend of large-scale morphometric datasets being collected by biological researchers, such as repositories consisting of thousands of images (Edlund et al., 2021; van der Linde et al., 2018). Traditionally, however, morphometric information is still extracted through manual annotation, which is time-consuming and labour-intensive. Furthermore, inter- and intra-observer measurement errors in landmark placements can lead to inconsistencies and faulty estimates (von Cramon-Taubadel et al., 2007).

Recent advances in artificial intelligence (AI) have opened up vast possibilities in the field of biology, including the potential to increase the scale, complexity, accuracy, efficiency, and reproducibility of phenotypic data collection (Wang et al., 2023). Computer vision is a prominent AI-based technology that allows for automated landmark placement and morphometric data collection, reducing variability introduced by manual annotation and improving the consistency of measurements across large datasets, and significantly decreasing processing time. Machine learning (ML) algorithms and methods common in computer science facilitate quick and accurate morphometric landmarking for complex dataset analyses that were previously impractical (He et al., 2024).

Given that machine learning-based phenotyping pipelines and infrastructure available to date are implemented in Python (e.g., Porto and Voje (2020)), a sufficient level of Python coding and machine learning knowledge is required for biologists to use these resources. To remove this skillset barrier and increase the accessibility of machine learning-based landmarking of photos to non-experts, a graphical user interface (GUI) can be an important attribute of such a research tool. Additionally, a free and open-source software has the benefit of being available to all researchers regardless of financial resources. Here, we present HusMorph, a free stand-alone and open-source machine learning application with a graphical user interface, automation of functions, and other user-friendly features to extract morphometric measurements from images. This stand-alone application runs on both Windows and Mac computers and makes it possible for anyone in the biological field to automate high-throughput morphometric analyses.

Materials and methods

We developed HusMorph as an accessible application with a user-friendly graphical user interface based on Python code. This interface allows users to quickly place and view landmarks on images with minimal user input (i.e., number of button presses). It also enables automated landmark placements using computer vision and machine learning techniques, primarily leveraging OpenCV 4.10.0.84, dlib 19.24.6, Optuna 4.1.0 and matplotlib 3.9.3 libraries. The interface is designed with intuitive and interactive elements (e.g. buttons and drop-down menus). Furthermore, the program is packaged as an executable file, allowing users to run it directly without needing to set up a Python environment or install additional dependencies, making it easy to use without requiring any coding knowledge. The application can be downloaded free from charge through GitHub (https://github.com/HenHus/Husmorph), and specific user instructions are available on the same web page. In the following sections, we provide a detailed explanation of the key features of HusMorph, and evaluate its performance on different datasets to validate the method and demonstrate its reliability. We tested the software on photos of fishes, although it should perform equally well on many other organisms or biological structures.

Image requirements

The machine learning training is easier if the images are optimised for the process. A homogenous background with a colour that is distinct from that of the organism is advantageous. We have mainly tested the process using zebrafish (*Danio rerio*) placed on their ventral side against a solid white background (Fig. 1A). However, machine learning can also work well with images containing slightly more complex backgrounds (Fig. 1B). Additionally, the method employed in the training process is sensitive to rotation, flipping, and scaling, indicating that better image standardization leads to improved results.

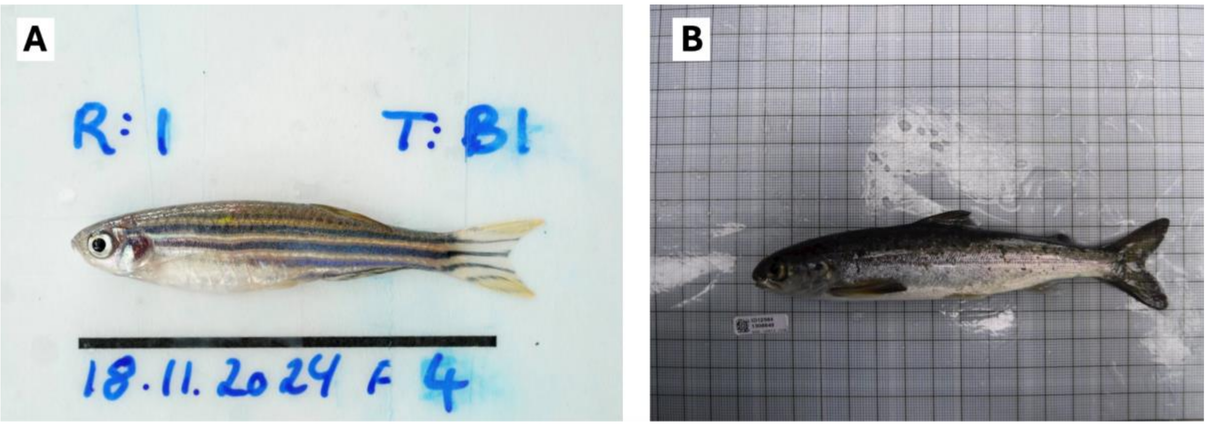

Figure 1. Examples of fish images suitable for machine learning training. (A) Zebrafish (*Danio rerio*) against a white background with few other image elements. The black scale bar is for length calibration. (B) Atlantic salmon (*Salmo salar*) against a slightly more complex background with light reflection, but with sufficient contrast from the organism of interest to allow for automation of landmarking.

As the machine learning training is computationally intensive, very high-resolution images require too much hardware without necessarily benefiting accuracy in landmark placements. Therefore, we recommend downscaling the images to match the processing power of the computer hardware available prior to training. We have used images around 1000 to 2000 pixels on the long side, and recommend this as the highest resolution to be used. This image resolution is amenable to training large sets (i.e., hundreds) of images on regular desktop computers and modern laptops. Very fast stationary computers (e.g., server clusters, workstations, gaming computers) may handle training sets with higher resolution images.

The number of images needed for successful training depends on the complexity and variability of the dataset, as well as the desired accuracy. Generally, a larger number of photos improves the accuracy of automated landmarking by increasing the diversity of training samples. To determine the optimal balance, we tested different training set sizes to identify the point where accuracy becomes sufficient and additional training data yield minimal improvement.

Computer hardware requirements

Training of machine learning algorithms is a computationally demanding process that requires both high-performance hardware and/or extended training duration. The accuracy of landmark predictions is highly dependent on the size of training datasets, with larger dataset and longer training times, generally leading to better results. To ensure efficient training, we primarily used a high-performance desktop computer with a high performance processor (Apple Mac Studio M2 Ultra). Additionally, we tested the training on various MacBook and Windows laptops. On modern laptops, a session of machine learning training can take around one to two days of continuous training, depending on factors such as dataset size, image resolution, number of machine learning trials, and hardware specifications. It is important to note that dedicating the whole processor to training may become hot and slow to perform other processes. However, machine learning training is a one-time process. Once the machine learning training is complete, it becomes substantially less computationally demanding, allowing it to place landmarks on large datasets of new images within seconds, even on standard laptops.

Automation of the machine learning process

To eliminate the need for users to possess an in-depth understanding of machine learning, the training parameters that affect the final outcome are not manually adjusted. Instead, 9 parameters are automatically optimized using the Optuna library—an open-source hyperparameter optimization framework that employs advanced algorithms to efficiently search for optimal parameter configurations (Akiba et al., 2019). Based on the training and testing conducted in this study, we have established a range for each parameter that performed well. Within these defined ranges, Optuna autonomously explores the optimal combination, avoiding the exhaustive search required by traditional methods like grid search—especially when dealing with rational-valued parameters. This streamlined approach reduces computational time and enhances overall efficiency.

Setting and viewing landmarks

One of the key features of the HusMorph app is its ability to efficiently place landmarks on images, providing a workflow similar to commonly used tools like ImageJ. It simplifies batch processing, as it saves landmarks in an XML format for seamless downstream use. While it can act as a stand-alone feature for quickly placing landmarks on smaller batches of images, its primary role is as a preprocessing step for training machine learning models.

Landmarks are represented as coordinate points on the image, with the origin located in the upper-left corner. The coordinate data are saved directly in an XML format, ensuring compatibility with machine learning workflows. Users can revisit and view the landmarks at any time, as the saved XML file contains all the necessary information for displaying the landmarks. This feature is particularly helpful for visually assessing the performance of machine learning models. By displaying the landmarks on the images, it provides a clear, intuitive way to evaluate accuracy - something that can be challenging to grasp from numeric measures like pixel deviation or percentage values alone.

Predicting landmarks

In this context, a *trial* represents a specific combination of the nine parameters used to train and test the machine learning model, while a *study* is a collection of trials aimed at identifying and saving the best-performing trial. Through the graphical user interface, users can specify the number of trials to include in a study, allowing them to balance the likelihood of finding a high-performing trial and minimizing the total time required to achieve the desired accuracy.

The tool automatically splits the dataset and performs a 5-fold cross-validation for each trial. The mean deviation across the folds is calculated and compared with other trials within the study to determine the most accurate configuration. The final selected model is derived from the best-performing fold from the trial with the lowest mean deviation, ensuring optimal accuracy.

Landmark accuracy

To quantify the accuracy in placement of landmarks, we measured the Euclidean distance in pixels between each landmark's true location and that predicted by the model, which is then normalized against the diagonal pixel distance of the total image. This is a different measure for accuracy compared to the deviation relative to the total length of the biological structure. The percentage error of the diagonal reflects how well the model handles the input image itself. Meanwhile, although the error normalized by the length of the biological structure is a conventional accuracy metric in the field, it fails to take into consideration the extent that the structure fills up the image frame. Indeed, an image of a small structure surrounded by a large amount of background is predisposed to have a lower deviation in pixels relative to the length of the structure and hence, higher accuracy, compared to an image of a large structure filling the frame. This puts the latter type of image in an unfair disadvantage when evaluating accuracy, because one pixel corresponds to a smaller physical distance. Therefore, by setting deviations as a percentage of the image diagonal - the biggest error possible - allow for an unbiased comparison of the prediction errors across different datasets and species.

To further validate the results, we will compare the model's performance with intra-observer error. Intra-observer error quantifies the variability in landmark placement when the same human expert annotates the same images on different occasions, serving as a baseline for manual accuracy. By comparing the model's deviation metrics to this error, we can assess whether the automated predictions achieve a level of consistency comparable to that of an expert. This evaluation helps test if the model's performance falls within an acceptable range of accuracy, examining its robustness and practical applicability for high-throughput morphometric analyses where reproducibility is critical.

Scale bar detection

The HusMorph machine learning application includes a convenient extension for automatically detecting and measuring scale bars in images. This feature provides an efficient way to convert pixel distances into more meaningful units such as millimetres. Images first undergo automatic pre-processing to enhance line features, after which the Hough Line Transform is applied to detect lines based on user-defined parameters for threshold, minimum line length, and maximum line gap. With minor coding adjustments tailored to the specific scale bar, the tool can be customized to ensure accurate calibration across all images. This functionality is valuable because even small movements of the camera or subject positions can cause variations in scale from image to image.

Testing dataset size and accuracy

To evaluate the performance of the application, we conducted tests using image datasets of varying sizes: 50, 100, 250, and 500 images. All images were standardized at a medium resolution of 1800x1200 pixels to analyse the relationship between dataset size and prediction accuracy. The images were randomly selected from a pool of 1,900 standardized images of zebrafish. Each image featured labelled zebrafish on a white background, all oriented in the same direction and positioned relatively horizontally with an extended caudal fin (Fig. 2).

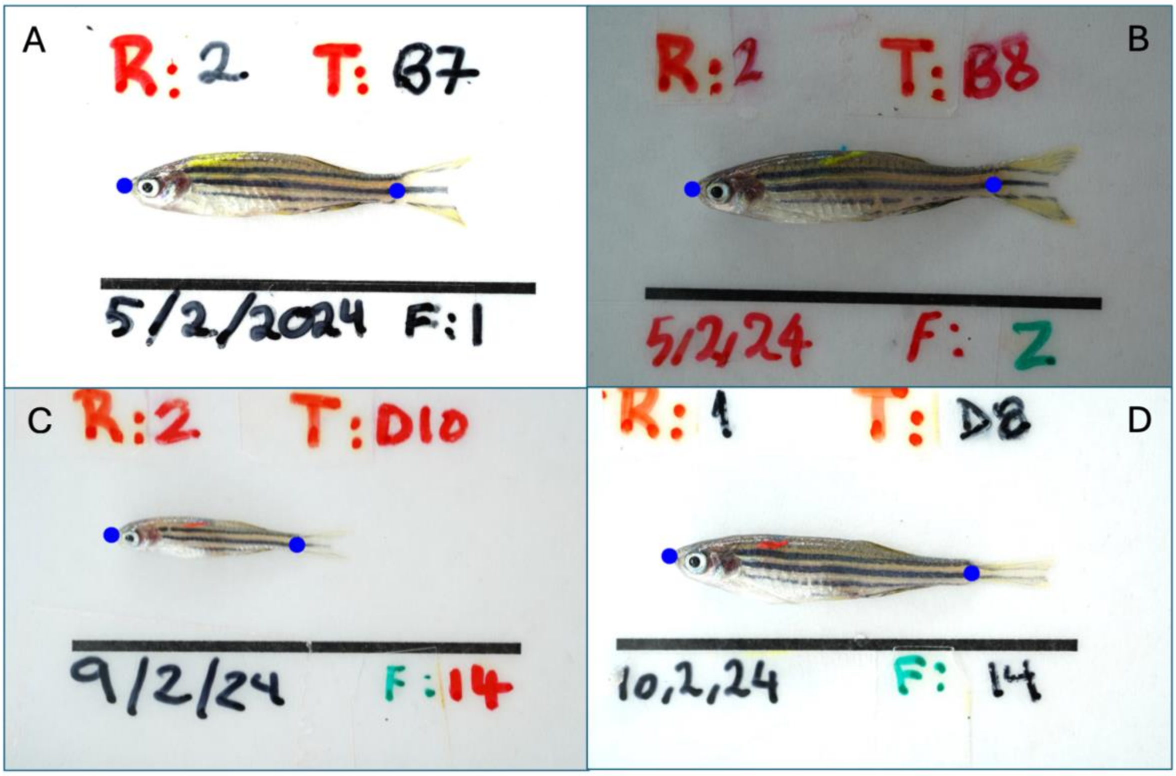

Figure 2: A selection of zebrafish (*Danio rerio*) images illustrating the variability within the landmark dataset used to evaluate app performance. (A) Slightly overexposed image, (B) highly underexposed image, (C) slightly blurred image featuring a very small zebrafish, and (D) zebrafish with an irregular, unspread caudal fin. The blue landmarks at the tip of the snout and the hypural joint of the caudal peduncle were used for determining standard length. The black scale bar below the zebrafish facilitates the conversion of pixel measurements to millimetres.

The dataset introduced variability, including being gathered by multiple people, two different cameras, differences in zebrafish size (ranging from 10 to 28 mm in standard length), variations in image clarity (e.g., focus and blur) and changes in lighting and exposure. This variability enabled evaluating the application robustness under varying image qualities and conditions in a standardized dataset.

Each image included two landmarks: one at the tip of the snout and another at the point where the caudal fin begins. Figure 2 illustrates these placements (blue filled circles). The two landmarks entailed different levels of detection difficulty. The snout landmark, located at the edge of the background, was easier to identify, while the caudal peduncle landmark, positioned at the hypural joint, was more challenging to detect. For clarity, we will refer to the landmarks as “simple” and “complex”, respectively, and intra-observer error as “human” error.

Testing a field dataset

To further assess the robustness of the landmark prediction, we conducted tests using hand-held field images of Atlantic salmon parr and smolts (Moccetti et al., 2024), which had minimal standardization (Fig. 1B). A total of 261 images with a resolution of 1200x900 pixels were analysed, with landmarks placed at the tip of the snout, the base of the caudal fin, the fork of the tail, and the base of the dorsal fin (Fig. 3). After completing the training process, an additional set of 29 previously unseen images was used for final validation to further evaluate the model's performance.

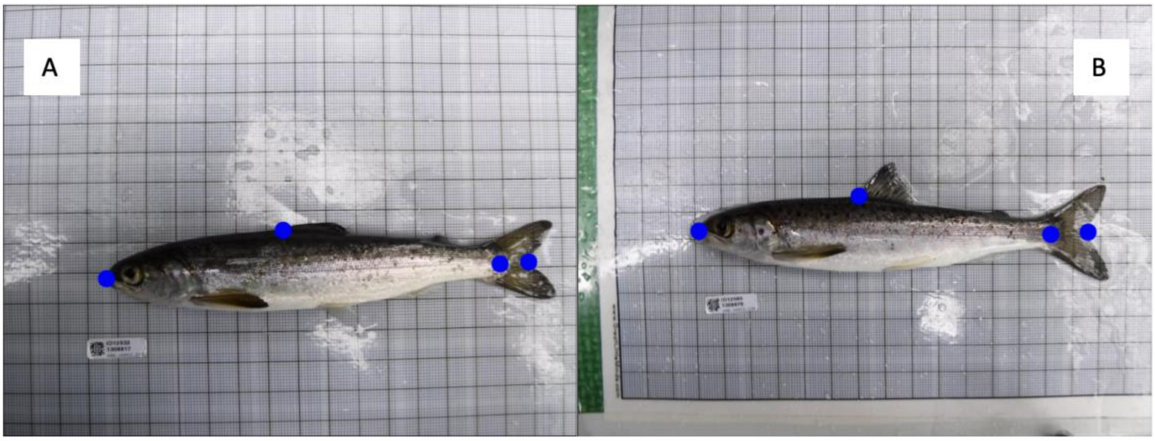

Figure 3: Subsample of the images of the Atlantic salmon used in experiment 2. Each image shows 4 blue dots, representing the 4 landmarks used; at tip of the snout, at the hypural joint of the caudal peduncle, at the fork of the tail fin and at the base of the dorsal fin.

Results

Image dataset size and variability

The accuracy was measured by measuring the performance on 1548 unseen images of zebrafish, and the accuracy was compared with intra observer error where the same person manually placed landmarks on the same dataset 9 months apart.

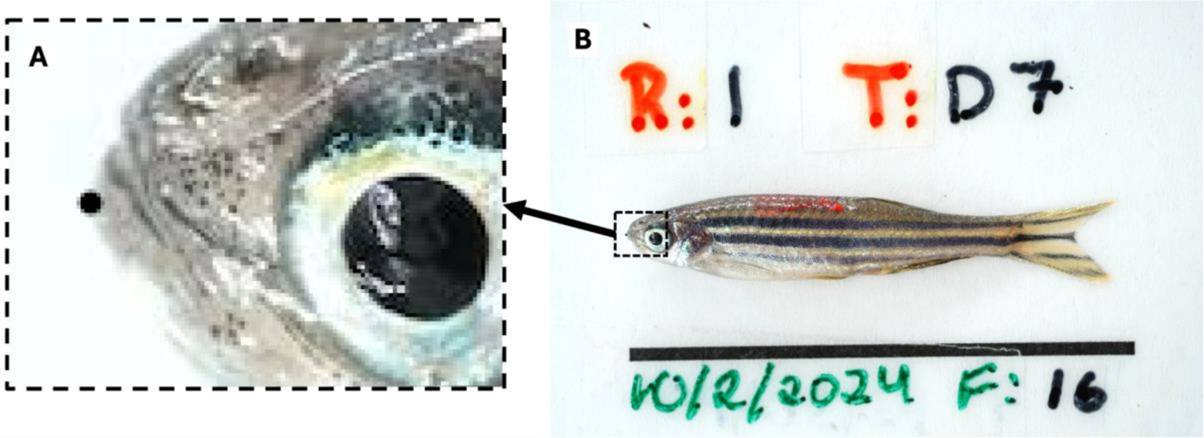

Figure 4: Visual representation of the accuracy of automated HusMorph landmarks. Image A is an enlargement of the full image shown in image B. The black circle in image A illustrates a 5-pixel deviation—a level of accuracy achieved after training on 500 images across 200 trials.

All models performed with accuracy <1% of image diagonal on the unseen images. The results showed that the dataset of 50 images has the highest mean error both for the simple landmark and the complex landmark (0.76% and 0.92% of image diagonal, or 16 and 20 pixels), and drastically dropped to dataset size of 100 (0.54% and 0.64% of image diagonal, or 12 and 13 pixels), before it dropped further for size 250 (0.24% and 0.37% of image diagonal, or 5 and 8 pixels), and stabilized towards the dataset of size 500 (0.23% and 0.32% of image diagonal, or 5 and 7 pixels). The simple landmark had overall less error compared to the complex error (Fig. 5). When compared to intra-observer error—0.19% of image diagonal (4 pixels) for the simple landmark and 0.28% of image diagonal (6 pixels) for the complex landmark—the model trained on 500 images was only 0.04% worse for both landmarks, approaching the theoretical maximum performance expected from a machine learning model.

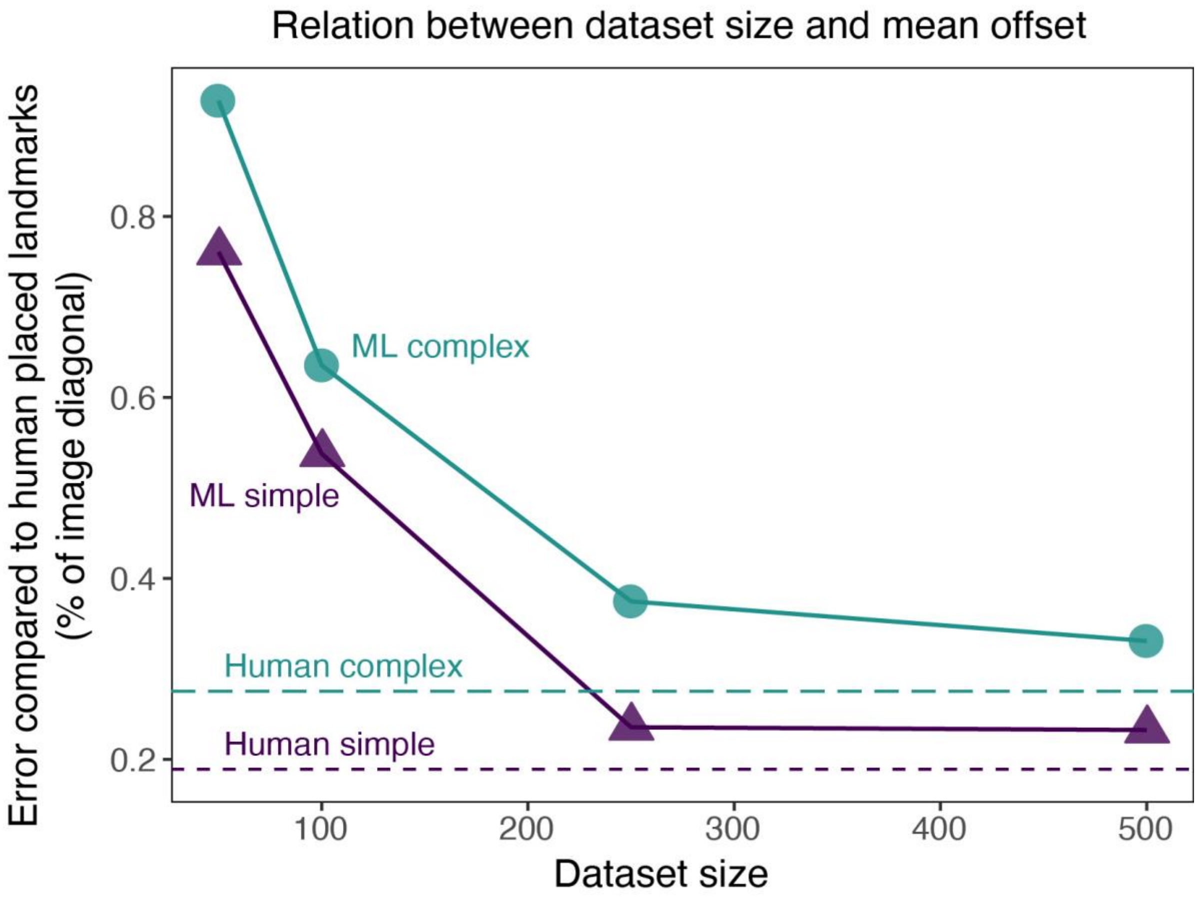

Figure 5: Comparison of the placement error between intra-observer error, marked as ‘human’, and automated machine learning landmark placements, marked as ‘ML’, as a function of dataset size. Two types of landmarks are tested, the simple (blue triangles) shows high accuracy for both human and machine learning landmarking, while the more difficult “complex” landmarks are overall more difficult to place.

Of the 200 trials conducted for each dataset, the dataset of size 50 exhibited significant improvements in the early trials but reached a performance plateau after trial 60. In contrast, the dataset of size 100 demonstrated a more gradual improvement throughout the study. Notably, the datasets comprising 250 and 500 images outperformed the smaller datasets, achieving high accuracy within the first 25 trials, with only marginal gains thereafter (Fig. 6A).

Figure 6B illustrates the probability of achieving various levels of deviation. Steeper curves indicate a higher likelihood of obtaining a model with low deviation. The figure demonstrates that datasets of 250, and 500 images have a higher than 80% chance of finding a model with less than 1% relative deviation. Sets of 100 images have around a 70% chance whereas a dataset of 50 images has a substantially lower chance of reaching the same level of accuracy. Notably, the small difference in performance between the 250- and 500-image datasets suggests that a dataset of around 250 images may offer the best balance between computational effort and accuracy.

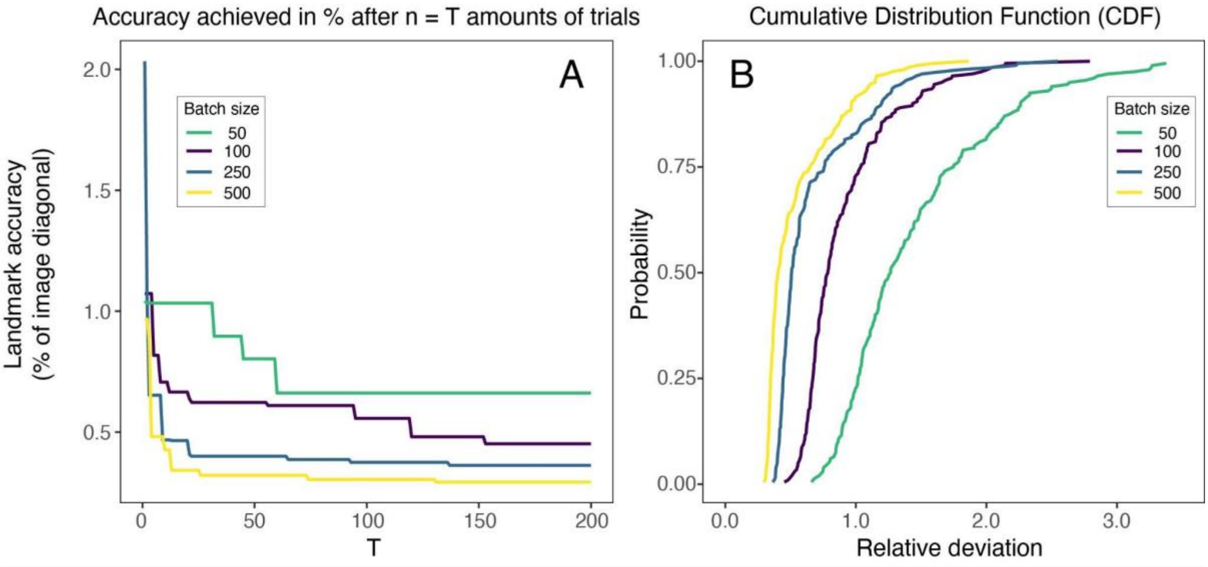

Figure 6. Machine learning landmark placement accuracy is dependent on the number of images (batch size) and the number of trials (x-axis) (A). Cumulative distribution function depicts the probability of reaching a certain deviation relative to the image diagonal using various image batch sizes (B). Image shows that larger training datasets have a higher chance of high accuracy.

Testing a field dataset

Based on the evolution of the study from experiment 1, we decided to perform a study of 100 trials, since the performance seemed to stabilize after this point. On the 29 unseen images, the trained model achieved a mean accuracy of 0.40% of image diagonal (6 pixels) per landmark. Intra-observer error was measured on the same 29 images with 1 day between the samplings, with a mean deviation of 0.21% of image diagonal (3 pixels).

In a related experiment, we were able to train a model using 1,935 images of zebrafish, achieving an accuracy of 0.18% of the image diagonal by following the same procedure provided by the application. To further assess the model's accuracy, a visual inspection of 6,966 predicted landmarks across 3,483 previously unseen images was conducted. The results indicated that 99.5% of the predicted landmarks were visually indistinguishable from those that would have been placed manually by an expert.

Discussion

Here, we show that our machine learning based HusMorph application can automatically place landmarks with a high accuracy on par with human experts. Using training sets of 100 to 250 photos, the application achieves a sufficient level of accuracy (5-14 pixels of deviation) from the optimal landmark location on high-definition images) for most biological purposes. A larger training dataset (more images) can further enhance accuracy (to ~6 pixels deviation), although it requires a higher computational effort (i.e., longer training time). We also show that bigger datasets have a very high chance of generating a well performing model that achieves good accuracy. It is generally not worthwhile to apply machine learning to smaller datasets (less than 100 images), as the accuracy was poor with training on 50 images. For smaller projects we therefore recommend placing the landmarks manually. However, beyond a dataset size threshold of around 100 images HusMorph offers a powerful solution to produce high-quality landmarks and automate morphometric data collection.

As expected, we found that the machine learning models detected simple (high contrast) landmarks more easily than complex (low contrast) landmarks regardless of dataset size. This suggests that the difficulty of landmark recognition (i.e., contrast to background and complexity of the landmarks) are crucial factors determining error rates and the level of accuracy. We demonstrate this when testing HusMorph to detect both fork length and standard length in fishes. Both measures use the tip of the snout as the anterior landmark; however, fork length uses the anterior end of the middle caudal fin rays, whereas standard length uses the hypural joint of the caudal peduncle. There is usually a clear contrast between caudal fin and background (Fig. 1), while differences in colouration and structure between caudal peduncle and caudal fin are much more subtle. Consequently, standard length landmarks are more difficult features to predict with machine learning. We should note, though, that it is common for human observers to have a similar difficulty in placing standard length landmarks with high accuracy (Carlander & Smith, 1945).

A major advantage of automating morphometric data collection from images is the reduction of systematic biases and errors inherent in human observation (Holman et al., 2015; Marsh & Hanlon, 2007). For instance, human observers that have preconceived expectations of results may subconsciously skew landmark placements in favour of this outcome. Further, the location and accuracy of landmarks placed by observers may drift over time, especially when morphometricians spend many days to collect a dataset, which causes disparity between measurements made across a timespan. Finally, when multiple annotators are involved, there is a risk of increased data variance or inter-observer errors. Therefore, an automated landmarking approach is preferable in many contexts (Clark et al., 2020; Clements JC et al., 2022; Jutfelt et al., 2017). Our machine learning based HusMorph application effectively achieves such automation, minimizes manual annotation errors and eliminates human biases.

The built-in machine learning procedure of HusMorph is well-established in the computer vision field. More advanced variants to this approach also exist—for example, those that combine object detection with landmark prediction (Porto & Voje, 2020) or 3D-pose estimation with tracking (Mathis et al., 2018). However, in our view, users are still required to have prior expertise in coding or machine learning to successfully build and use such machine learning pipelines. We specifically designed HusMorph to be accessible to users without specialized technical knowledge, by automating the parameter optimization of the machine learning model and packing the method in a user-friendly application with an intuitive graphical user interface. This application is not intended to replace experts in the machine learning field; rather, it serves as a supplementary tool, enabling high-throughput analysis without the need for specialized technical knowledge. While expert users can also benefit from its capabilities, its primary focus is to support those without advanced expertise in the field. By integrating complex machine learning processes into a program with an easy-to-use graphical user interface, HusMorph facilitates automation in phenotyping, allowing researchers in the morphometric community to quickly assess biological structures in images at a large scale regardless of technical background.

The dlib library used in our machine learning software differs from typical CNN-based predictors, which typically train on large datasets containing thousands of input images and are harder to fine tune (Xue et al., 2021). Predictive models developed this way are generally more robust, exhibiting higher tolerance to image transformations such as rotation, scaling, and flipping, which can be weaknesses of our dlib library-based approach. However, most research studies in biology neither require nor have access to the resources that would support that level of versatility. Therefore, for moderately sized (i.e., hundreds of images) and standardised datasets that are common in biological studies, the dlib library which HusMorph uses offers a great compromise between tolerance to variation in image quality and machine learning effort.

Table 1: Recommended dataset specifications and built-in machine learning procedure steps and features used for automated landmarking with the HusMorph application.

Component	**Description**
Input image recommendations	
Quality	Standardized rotation, flipping and scaling
Number	≥100
Resolution	≤ 2 megapixels
Background	Distinct from object, ideally homogeneous
Other	Consistent lighting and focus
Built-in machine learning procedure steps and features	
Manual landmarking in the training set	Ground truth in the training dataset is established through expert manual landmark placements.
Machine learning model training and testing	Machine learning models are generated from the set of images with the manually placed landmarks.
Automated landmarking and visual confirmation	The model automatically places predicted landmarks on unseen images, which users can review to confirm accuracy.
Automated scale bar detection	Optional automated scale bar detection script converts pixels to length units for each image in batch mode.
Export of predicted landmarking results	Predicted landmark results can be conveniently exported in CSV format.

Conclusions

The open-source HusMorph application allows researchers to quickly automate morphometric data collection from large sets of images. The program has an easy-to-use graphical interface and therefore allows anyone to employ the power of machine learning to their morphometric research questions, with a minimal learning curve. We show that the machine learning model can reach similar accuracy to human observers through training on a few hundred images over a day or two on a normal PC or Mac computer. We therefore hope that the HusMorph application will benefit researchers by increasing speed and ease of data collection, while at the same time promoting accuracy through reductions in biases and variation from human observation.

Acknowledgements

We wish to thank Arthur Porto and Christophe Pélabon for valuable input during early discussions about automated image analysis using machine learning and working with the ML-morph pipeline. We also thank Paolo Moccetti and all student assistants in our research group for contributing or collecting fish images for model training and validation of HusMorph. Finally, we are grateful to Pierre Renaud for valuable feedback on the user experience of HusMorph during application development.

Author contributions

H.H.K., R.H.J.L. and F.J. conceived the ideas and designed methodology; all authors contributed to data collection; H.H.K. conducted the formal analysis; H.H.K., M.M. and F.J. were responsible for visualization; H.H.K., R.H.J.L. and F.J. led the writing of the manuscript (original draft). All authors contributed critically to the review and editing of the manuscript, and F.J. was responsible for funding acquisition.

Conflicts of Interest

The authors have no conflicts of interest to declare.

Funding

This work was supported by a European Research Council Consolidator grant to F.J. (number 101003026-CLIMEVOLVE).

Data Availability

The HusMorph application, user instructions, scripts, and image files can all be found on GitHub (https://github.com/HenHus/Husmorph).

References

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). *Optuna: A Next-generation Hyperparameter Optimization Framework* Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Anchorage, AK, USA. https://doi.org/10.1145/3292500.3330701

Carlander, K. D., & Smith, L. L. (1945). Some Factors to Consider in the Choice between Standard, Fork, or Total Lengths in Fishery Investigations. *Copeia*, *1945*(1), 7-12. https://doi.org/10.2307/1438165

Clark, T. D., Raby, G. D., Roche, D. G., Binning, S. A., Speers-Roesch, B., Jutfelt, F., & Sundin, J. (2020). Ocean acidification does not impair the behaviour of coral reef fishes. *Nature*, *577*(7790), 370-375. https://doi.org/10.1038/s41586-019-1903-y

Clements JC, Sundin J, Clark T.D, & F, J. (2022). Meta-analysis reveals an extreme “decline effect” in the impacts of ocean acidification on fish behavior. *PLoS Biol 20(2): e3001511.* https://doi.org/e3001511

Edlund, C., Jackson, T. R., Khalid, N., Bevan, N., Dale, T., Dengel, A., Ahmed, S., Trygg, J., & Sjögren, R. (2021). LIVECell—A large-scale dataset for label-free live cell segmentation. *Nature Methods*, *18*(9), 1038-1045. https://doi.org/10.1038/s41592-021-01249-6

He, Y., Mulqueeney, J. M., Watt, E. C., Salili-James, A., Barber, N. S., Camaiti, M., Hunt, E. S. E., Kippax-Chui, O., Knapp, A., Lanzetti, A., Rangel-de Lázaro, G., McMinn, J. K., Minus, J., Mohan, A. V., Roberts, L. E., Adhami, D., Grisan, E., Gu, Q., Herridge, V., . . . Goswami, A. (2024). Opportunities and Challenges in Applying AI to Evolutionary Morphology. *Integrative Organismal Biology*, *6*(1). https://doi.org/10.1093/iob/obae036

Holman, L., Head, M. L., Lanfear, R., & Jennions, M. D. (2015). Evidence of experimental bias in the life sciences: why we need blind data recording. *PLoS biology*, *13*(7), e1002190.

Jutfelt, F., Sundin, J., Raby, G. D., Krång, A. S., & Clark, T. D. (2017). Two-current choice flumes for testing avoidance and preference in aquatic animals. *Methods in Ecology and Evolution*, *8*(3), 379-390.

Klingenberg, C. P. (2010). Evolution and development of shape: integrating quantitative approaches. *Nature Reviews Genetics*, *11*(9), 623-635.

Marsh, D. M., & Hanlon, T. J. (2007). Seeing what we want to see: Confirmation bias in animal behavior research. *Ethology*, *113*(11), 1089-1098.

Mathis, A., Mamidanna, P., Cury, K. M., Abe, T., Murthy, V. N., Mathis, M. W., & Bethge, M. (2018). DeepLabCut: markerless pose estimation of user-defined body parts with deep learning. *Nature Neuroscience*, *21*(9), 1281-1289. https://doi.org/10.1038/s41593-018-0209-y

Moccetti, P., Rodger, J. R., Bolland, J. D., Kaiser-Wilks, P., Smith, R., Nunn, A. D., Adams, C. E., Bright, J. A., Honkanen, H. M., Lothian, A. J., Newton, M., & Joyce, D. (2024). The reproducibility of geometric morphometric analyses on live fish. https://doi.org/10.17605/OSF.IO/ZVU5P

Porto, A., & Voje, K. L. (2020). ML-morph: A fast, accurate and general approach for automated detection and landmarking of biological structures in images. *Methods in Ecology and Evolution*, *11*(4), 500-512.

Rohlf, F. J., & Marcus, L. F. (1993). A revolution morphometrics. *Trends in ecology & evolution*, *8*(4), 129-132.

Rowiński, P. K., Mateos-Gonzalez, F., Sandblom, E., Jutfelt, F., Ekström, A., & Sundström, L. (2015). Warming alters the body shape of European perch Perca fluviatilis. *Journal of fish biology*, *87*(5), 1234-1247.

van der Linde, S., Suz, L. M., Orme, C. D. L., Cox, F., Andreae, H., Asi, E., Atkinson, B., Benham, S., Carroll, C., Cools, N., De Vos, B., Dietrich, H.-P., Eichhorn, J., Gehrmann, J., Grebenc, T., Gweon, H. S., Hansen, K., Jacob, F., Kristöfel, F., . . . Bidartondo, M. I. (2018). Environment and host as large-scale controls of ectomycorrhizal fungi. *Nature*, *558*(7709), 243-248. https://doi.org/10.1038/s41586-018-0189-9

von Cramon-Taubadel, N., Frazier, B. C., & Lahr, M. M. (2007). The problem of assessing landmark error in geometric morphometrics: theory, methods, and modifications. *American Journal of Physical Anthropology: The Official Publication of the American Association of Physical Anthropologists*, *134*(1), 24-35.

Wang, H., Fu, T., Du, Y., Gao, W., Huang, K., Liu, Z., Chandak, P., Liu, S., Van Katwyk, P., Deac, A., Anandkumar, A., Bergen, K., Gomes, C. P., Ho, S., Kohli, P., Lasenby, J., Leskovec, J., Liu, T.-Y., Manrai, A., . . . Zitnik, M. (2023). Scientific discovery in the age of artificial intelligence. *Nature*, *620*(7972), 47-60. https://doi.org/10.1038/s41586-023-06221-2

Xue, H., Artico, J., Fontana, M., Moon, J. C., Davies, R. H., & Kellman, P. (2021). Landmark detection in cardiac MRI by using a convolutional neural network. *Radiology: Artificial Intelligence*, *3*(5), e200197.