Convening Expert Taxonomists to Build Image Libraries for Training Automated Classifiers

Kasia M. Kenitz, Eric C. Orenstein, Clarissa R. Anderson, Alexander J. Barth, Christian Briseño-Avena, David A. Caron, Melissa L. Carter, Emily Eggleston, Peter J. S. Franks, James T. Fumo, Jules S. Jaffe, Kelsey A. McBeain, Anthony Odell, Kristi Seech, Rebecca Shipe, Jayme Smith, Darcy A. A. Taniguchi, Elizabeth L. Venrick, and Andrew D. Barton

Abstract

Digital imaging technologies are increasingly used to study life in the ocean. To deal with the large volume of image data collected over space and time, scientists employ various machine learning and deep learning algorithms to perform automated image classification. Training of classifiers requires a large number of expertly curated sets of images, a time-consuming process that requires taxonomic knowledge and understanding of the local ecosystem. The creation of these labeled training sets is the critical bottleneck for building skillful automated classifiers. Here, we discuss how we overcame this barrier by leveraging taxonomic knowledge from a group of specialists in a workshop setting and suggest best practices for effectively organizing image annotation efforts. In our experience, this 2 day workshop proved very insightful and facilitated classification of over 4 years of plankton images obtained at Scripps Pier (La Jolla, CA), focusing on diatoms and dinoflagellates. We highlight the importance of facilitating a dialog between taxonomists and engineers to better integrate ecological goals with computational constraints, and encourage continuous involvement of taxonomic experts for successful implementation of automated classifiers.

What bottleneck?!

Imaging technologies are an increasingly common tool used to observe marine organisms (Sosik and Olson 2007; Lombard et al. 2019; Kenitz et al. 2023). They enable new sampling designs and strategies, and are capable of observing organisms at high rates in space and time (Irisson et al. 2022). Yet high-frequency sampling brings with it the new challenge of analyzing a large volume of images. Over the past 10 yr, scientists have begun to leverage advances in machine learning and pattern recognition to efficiently classify and quantify image data (Sosik and Olson 2007; Orenstein et al. 2020*a*). Most researchers use some type of supervised machine learning, in which a computer classifier is taught to recognize patterns in images by tuning an algorithm with an expertly curated set of images.

The collection of high-quality annotated datasets for training is a major bottleneck for projects seeking to use machine learning tools (Schoening et al. 2016). While there are many resources that guide beginners through application of machine and deep learning tools, there is less attention given to curation of a suitable training dataset that classification algorithms rely upon. Image annotation requires careful inspection of images and manual classification into appropriate categories based upon domain expertise (domain expertise could be, e.g., a plankton taxonomist with knowledge of communities in the California Current region). This is challenging work, requiring thoughtful dataset design and many hours of effort by trained experts. Most research efforts produce the labeled training set with a single expert or a small cohort of experts working together who, in many cases, may have limited understanding of the deep learning process. Some projects outsource collection of training data, either "gamifying" the procedure to attract volunteers, or paying private contractors. Each approach has merits and drawbacks; doing everything internally allows for greater consistency but can be time-consuming, especially when no funds are available to ensure a full-time commitment. Crowdsourcing can produce a training set quickly but is prone to error as the human annotators may not be specifically trained for the task (Irisson et al. 2022). Here we discuss soliciting taxonomic knowledge from a large group of specialists in a targeted, short-duration workshop setting as an effective and speedy method for creating and curating a large and complex training dataset.

In July 2019, we organized a focused, 2-d taxonomy and machine learning workshop at Scripps Institution of Oceanography, La Jolla, CA (Fig. 1). Our goal was to bring

FIG. 1. Photos of some (unfortunately not all!) participants of the Taxonomy Workshop at Scripps Institution of Oceanography, 9–10 July 2019.

together expert plankton taxonomists, ocean engineers, and machine learning specialists to facilitate the swift annotation of plankton images for training of automated classifiers and foster a community to assist with future efforts in validating the classifier output. The plankton images were collected by the Scripps Plankton Camera (SPC), located on Scripps Pier at a water depth of 3 m (Orenstein et al. 2020*b*). The camera images drifting particles in situ, capturing phytoplankton and zooplankton species, as well as detritus, sand, and other particles. Over the course of the workshop, 18,303 images in 35 classes were annotated, with the goal of training an automated classifier to track the abundance of these plankton and particle groups through time. These time series will have a broad range of applications, ranging from understanding the factors that influence harmful algal blooms (HABs) on short time scales, to linking anthropogenic changes in climate to ecosystems dynamics.

Here, we discuss our method for creating a comprehensive image dataset—in terms of both the number of labeled images and the structure of the dataset—elaborating on the image classification workflow and issues that inhibited efficient annotation. We provide concrete recommendations on how to design and build an image database for researchers without considerable prior knowledge or experience in building plankton image libraries. We show that the workshop setting can significantly accelerate designing and applying the image annotation workflow, and that just 2 d of concentrated effort can provide an insight into the outcome of automated classification and indicate the direction for further improvement. Although this project is focused on classification of underwater plankton images collected by the SPC, we believe that the lessons are widely applicable, regardless of the imaging system used or target organisms considered, and can be adapted to a virtual setting in the case that an in-person workshop is not possible. We conclude that sustained and significant progress in automated monitoring of planktonic communities in future efforts will be facilitated by implementing the lessons and the workflow discussed here (Box 1), and highlight the merits of a focused effort to dataset design and annotation collection in a workshop setting. In addition to the methodological suggestions, we highlight the need for continuous involvement of taxonomic experts to ensure successful implementation of automated classifiers.

The approach: How and why?

Our goal was to design an annotation workflow from scratch and build a large image database for training of automated classifiers. An important first step was to consider the instrument used for collecting underwater images, which in our case was the SPC system. Then, we followed with a series of scientific and engineering considerations necessary for designing a classification workflow and involving taxonomists in the classification design and image annotation.

SPC system

The SPC is a digital, in situ, darkfield, imaging system designed to sample organisms and particles in their aquatic setting without disturbing the ambient flow or interfering with imaged particles (Orenstein et al. 2020*b*). The system includes two microscope objectives (×0.5 and ×5) and captures images of particles that range between tens of micrometers to several centimeters in size. The system is comprised of three distinct parts: the in situ cameras, the database of images, and a front-end web interface. The SPC is a free-space system which means it uses no nets, filters, or pumps, and only captures images within ambient sea water. The underwater unit consists of two housings for each microscope, one with illumination optics and the other with the imaging system, separated by an uninterrupted volume (Fig. 2). After a full-frame image is acquired by the SPC, the image processing routine in the on-board processing unit segments bright objects from the dark background (Orenstein et al. 2020*b*). The web interface (http://spc.ucsd.edu/) was designed to allow registered users to browse data by object size and time, and to annotate imaged organisms with taxonomic or semantic labels with the ability to define new classes as needed. Since deployment in March 2015, the SPC has collected over 25 terabytes of data and over one billion images of individual organisms and particles.

Designing classification categories

We describe here an iterative approach for defining and refining classifier categories that incorporates aspects of the imaging system and the expertise of taxonomists about what is practical and tractable. An important first

BOX 1. Objectives in designing an automated classification workflow

1. Establish the *scientific goals* and research questions that the ecological data produced by the automated classifier will be used to answer. *Is the research focused on community or species-specific variability? Is there a need to include rare species?*
2. Determine the level of *taxonomic detail* required for image classification that addresses the scientific objectives. *Which taxa can be grouped into broader groups? Which taxa need to be considered at the species level?*
3. Consider *engineering limitations* when constructing a classification scheme. *Does the instrument have sufficient resolution to capture necessary morphological detail to enable clear distinction between closely related organisms? Which taxa are too challenging to be distinguished if imaged in suboptimal quality?*
4. In the case of a multi-class classification (in contrast to binary classification), carefully consider what constitute *noise classes*.
5. Select an appropriate *classification algorithm* and determine the number of images per class required for classifier training and successful implementation of the algorithm.
6. Ensure that images used for algorithm training are collected over a wide temporal range to facilitate collection of comprehensive training data that captures a wide *range of intra-species morphologies and across instrument states*.
7. *Consider the resources* when deciding on the number of classes, including time and expertise necessary for validation of the classifier output.

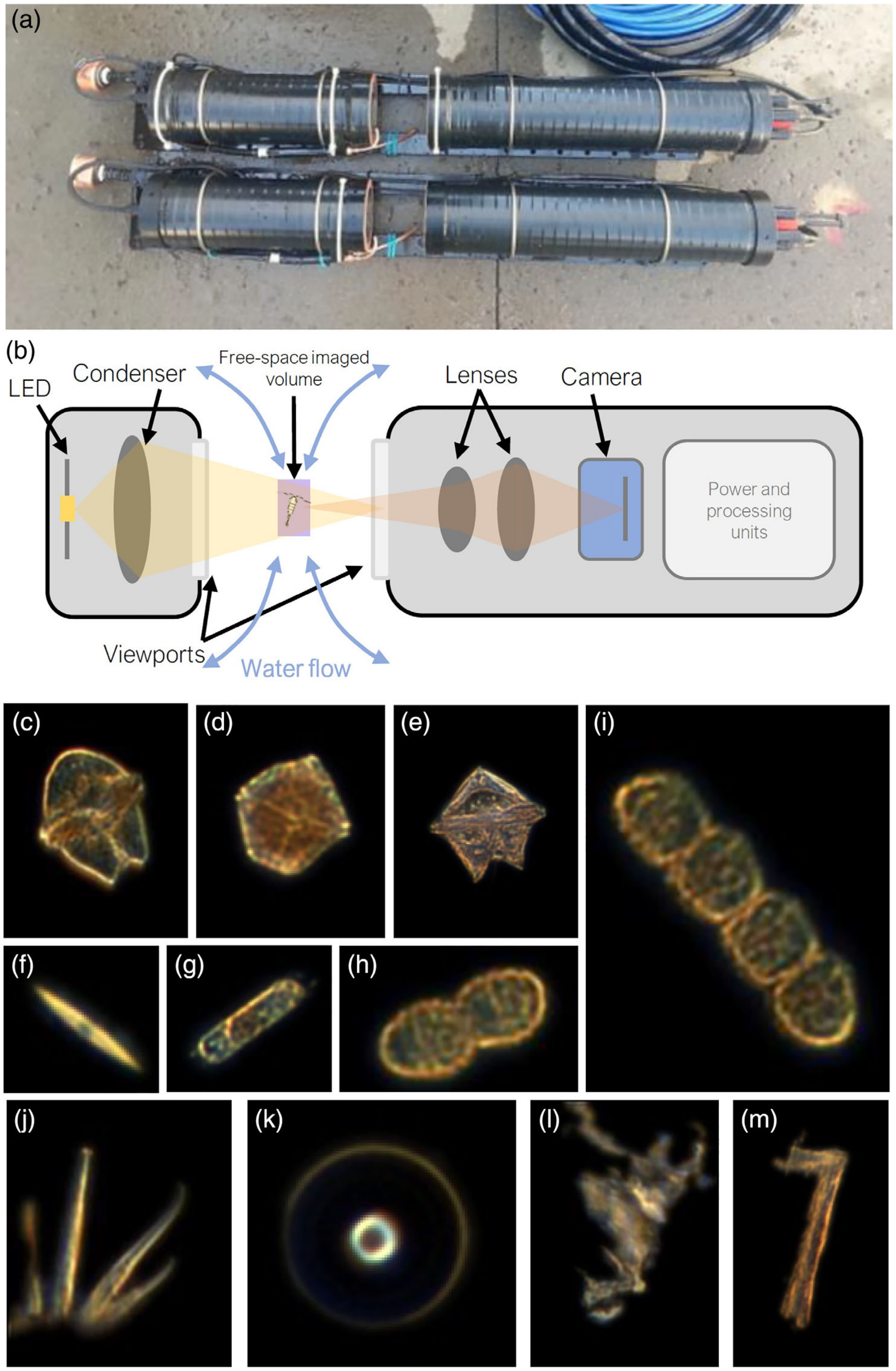

FIG. 2. On the left, (**a**) photo of the Scripps Plankton Camera System and (**b**) a schematic diagram of its design. On the right, example images captured by the camera and these include (**c**) *Akashiwo sanguinea*, (**d**) *Lingulodinium polyedra*, (**e**) *Protoperidinium* sp., (**f**) unidentified pennate and (**g**) centric diatoms, (**h**, **i**) *Margalefidinium* sp., and (**j–m**) example images sorted into distinct noise classes.

step when creating a classification scheme is identifying the scientific goals and objectives that the automated classifier will be used to fulfill. Establishing the scientific purpose is essential to defining the relevant classes for the automated classifier and the necessary level of taxonomic identification. A crucial second step is to combine the scientific requirements with technical capabilities of the instrument collecting the data—sample volume, image resolution, and so forth—to inform what can be observed and quantified.

Scientific objectives. The primary purpose of the SPC development and deployment is to complement ecological and environmental monitoring efforts conducted by the Southern California Coastal Ocean Observing System (SCCOOS) at Scripps Pier by providing high-temporal resolution information on the phytoplankton and zooplankton communities, focusing in particular on the HAB taxa. Hence, the automated classifier focuses on common genera (e.g., *Tripos*, *Chaetoceros*, *Protoperidinium*, or *Prorocentrum* spp.) as well as taxa that make irregular, yet ecologically significant appearances, in particular, harmful algae including *Pseudo-nitzschia* spp., *Alexandrium* spp., *Akashiwo sanguinea* (Fig. 2a), *Lingulodinium polyedra* (Fig. 2b), *Dinophysis* spp., and *Chattonella* spp. Ideally, the classifier should also enumerate the potential competitors (other phytoplankton, for example) and predators (microzooplankton, mesozooplankton) of the HAB taxa.

Instrument design constraints for taxonomic identification. The ability to observe these taxa using visual information is constrained by the ability to distinguish different species that share somewhat similar morphological features. For image data, the instrument's design constrains the detail that can be observed. With the SPC (Fig. 2), the key aspects are the pixel resolution, illumination source, free-space design and the continuous in situ deployment (Orenstein et al. 2020*b*).

For example, the SPC's pixel size limits the scale of structures that can be resolved, such as spines on some organisms that can be too thin to be captured by the imaging system (Orenstein et al. 2020*b*). Darkfield imaging enhances edge detection and preserves pigmentation information for translucent organisms, but may fail to capture intricate structural details that enable species-specific identification of some taxa. The free-space imaging design captures organisms in their natural state without disruption, however, this feature means that organisms are randomly oriented relative to the camera and can show up in or out of focus. Finally, ambient environmental conditions and biofouling during deployment can deteriorate image quality and limits image classification.

For the purpose of the workshop, we compiled a list of tentative phytoplankton and microzooplankton categories for automated classification with these constraints in mind. The set was then further discussed and refined based on the feedback from the experts during the workshop. For example, within the genus of *Tripos* (previously *Ceratium*), the species *T. furca*, *T. fusus*, and *T. falcatiformis* are the most commonly observed at Scripps Pier. *T. furca* composed its own individual class. But, upon consideration of the instrument limitations, *T. fusus* and *T. falcatiforme* were grouped into a single class because critical morphological features would often be indistinguishable to a human annotator when imaged at certain angles or in poor resolution. The remaining, less abundant *Tripos* species were grouped together into a single class as "Tripos_other." Note that species or genus names are not italicized here if used as a name of a classification category.

Instrument design also inhibited species-specific identification within the genus of *Chaetoceros*, where the inability to capture the intricate details of fine structures connecting cells in a chain made species-level distinctions difficult. For cells of sizes approaching the lower resolution limit of the imaging system, or when image quality was poor, class distinctions were made based on broad morphological differentiation, for example, "unknown pennate diatom" vs. "unknown centric diatom" (Fig. 2d,e). Our classification scheme also distinguished cells that could be observed in a solitary form as well as in chains (Fig. 2g). Therefore, where technical limitations prevented taxonomy-based classifications, semantic descriptors (e.g. "single cell," "chain," "spines") were used instead of the taxonomic identifiers. Overall, our methodological approach for developing classifier categories was iterative; we entered the workshop with a general concept of desired classifier categories, but revised and refined this with the help of taxonomic experts.

We focused on annotating and classifying organisms and particles in the size range 30–1000 μm, though the SPC captures larger and smaller organisms and particles. The lower size range was restricted by the pixel resolution of the camera that inhibits image clarity, and thus reliable classification of particles smaller than 30 μm. The upper limit was selected in consideration of the size of organisms of interest and the likelihood of collecting a sufficient number of large organisms that are rare relative to the sample volume.

Important considerations for the training dataset. When building an image database for automated classification it is important to account for environmental variability that not only affects the relative abundance but also the appearance of organisms and particles, both directly and indirectly. The appearance of target organisms may differ due to cell growth phase (i.e., cell division), feeding stage (of importance for mixotrophic or heterotrophic eukaryotes that were imaged after having engulfed their prey), and cell orientation in the imaged volume. Biofouling on the instrument may also impact the image quality. Together, these factors contribute to *dataset shift*, where the target data are different from the training data leading to lower-quality automated population estimates (González et al. 2017; Orenstein et al. 2020*a*). It is therefore important for the training dataset to contain images collected over a range of environmental and instrument maintenance scenarios to ensure that all recognizable states of the organisms are represented in the training set.

To incorporate temporal variability in the training set, it is possible to randomly select time periods for image annotation over the period of instrument deployment. If available, one can also leverage scientific literature or ecological data that captures local or regional variability in the plankton populations to inform the selection of specific time periods when target taxa were relatively abundant to ensure efficient image annotation. We used weekly monitoring reports for Scripps Pier collected by SCCOOS as part of the Harmful Algal Bloom Monitoring and Alert Program (HABMAP) (https://calhabmap.org) to inform the selection of key taxa for classification. We also used these data to identify time periods of elevated abundance of selected species, which we expected to coincide with periods when the camera would record increased occurrences of the corresponding classified taxa. To ensure the collection of sufficient training data, we chose to semi-randomize the selection of dates for image annotation by biasing toward high abundance days according to the monitoring data at Scripps Pier, instead of uniformly sampling across the time period of the instrument deployment. This approach is particularly useful to ensure collection of images of rare, but ecologically relevant, organisms that form infrequent and often short-lived blooms. This approach, however, is biased toward capturing organism morphologies *during* bloom conditions, which may lead to poor classifier performance during the non-bloom periods if the organism's morphology is significantly different. On the other hand, if the goal of the project is to identify blooms, poor classifier performance during non-blooms conditions may not be a critical flaw.

Image dataset builders must pay close attention to images of organisms or particles that are not of immediate interest to the project, relegating them to "noise." In our case, "noise" categories incorporate any nonliving or non-biological particles, such as sand, marine snow, or gas bubbles (Fig. 2h–k). These particles, while not of primary interest to our study of plankton, are morphologically distinct and should be treated as separate classes. Without accounting for noise categories, the classifier will distribute the images of such objects among taxonomic classes, negatively impacting classifier performance (Li and Vasconcelos 2020). The more specific the noise categories are, the better the classifier performance (Li and Vasconcelos 2020; Orenstein et al. 2020*a*).

Resource requirements for a successful classifier. It is important to consider the amount of human effort necessary to build an image dataset for training of the classifier, where the required volume of image data is determined by the number of resolved classes and the choice of the classification algorithm. In addition, human effort may be necessary for further validation of the classifier output, where the required labor increases linearly with the number of classes.

For automated image classification, we focused our efforts on the increasingly common deep convolutional neural networks (CNNs) that require expert annotated data for training purposes. The architecture of most CNNs calls for 1000s of example images per class to appropriately tune the algorithm (LeCun et al. 2015). In general, a deeper CNN with more layers will have greater representational power and greater ability to learn more complex patterns. Increasing the depth and thus the complexity of the network comes at the cost of requiring increasingly larger training datasets. As a consequence, a deeper network is not always the best choice due to resource limitations. Rather, the optimal architecture depends on the amount of data available for training. Techniques such as fine tuning (i.e., reusing lower-level feature extraction layers of the neural network) and data augmentation (i.e., creating new images from random affine transformations of the existing ones) can be effective when limited training data are available. The selection of a suitable classification architecture can be made by comparison of classifier performance across a range of different architectures, paying careful attention not to "overengineer" the network and risk overfitting. Since the selection of the optimal architecture was not the objective of the workshop, we have selected a classifier structure that was previously proven to work well on image data collected by the SPC (Kenitz et al. 2020; Orenstein et al. 2020*a*).

Recent studies highlight the need for human validation of the classifier output to ensure high-quality ecological data (Axler et al. 2020; Kenitz et al. 2020). Automated classifiers often assume that the properties captured in the training dataset reflect those observed in situ. The phenomenon of dataset shift, in which the data used for training do not reflect the day-to-day variability observed in situ, is pervasive in natural environments because the relative abundances and appearances of organisms are rarely static. Dataset shift degrades the performance of automated classifiers, but can be mitigated with a variety of automated or semi-supervised quantification methods (González et al. 2017; Orenstein et al. 2020*a*). Validation of the classifier output may require more resources and effort than collection of training images, and should be accounted for when designing a classification scheme.

Involving the experts: Workshop design

We invited taxonomists who specialize in ecosystem and HAB monitoring efforts on the U.S. West Coast, the majority coming from the state of California. A total of 17 scientists with taxonomic expertise assisted with annotation of plankton images collected by the SPC. The focused, 2-d workshop was designed to: (1) create opportunities for networking, sharing, and discussing new scientific ideas through inviting participants to present their research and participate in casual networking events; (2) familiarize taxonomic experts with the machine learning tools and their applications; (3) educate engineers about the desired scientific outcomes by facilitating discussions with taxonomists and showcasing their research interests; and (4) emphasize the importance of including taxonomic expertise in the pipeline as automated technologies increasingly come online.

To ensure collection of high-quality labeled data, it is important that the annotators understand key principles and limitations of the machine learning tools, and how the selection of images used for training affects the performance of machine classifiers. Prior to image annotation sessions, we offered a brief "crash course" on machine learning and computer vision aimed at familiarizing all annotators with machine learning concepts and highlighting important considerations when building an image dataset. We emphasized the need to include: (1) blurry or poor-resolution images, as long as the organism is still clearly identifiable and (2) images of organisms captured at suboptimal orientations or angles, as long as structural details are captured sufficiently to enable taxonomic identification. We thus encouraged annotators to label images when they were at least 75% confident in their taxonomic identification to ensure that the classifier had training data that encoded imperfect examples. With enough training data, the CNN should be able to learn the differences between each class in the context of such variability (LeCun et al. 2015). Otherwise, the algorithm will only learn to recognize appropriately oriented, in-focus images of a particular organism.

Prior to image annotation sessions, we held a discussion of the provisional list of classes for automated classification that was prepared prior to the workshop, and examined example images for each class. The discussion focused on the feasibility and limitations of species identification, especially when imaged at suboptimal angels, and the final list of annotation classes was formed based upon the group consensus and the available taxonomic expertise. This discussion was particularly important for establishing the classification scheme for chain-forming diatoms as the camera was often unable to capture the characteristic morphological features allowing for differentiation between chain-forming diatom taxa, enforcing the implementation of semantic class descriptors.

The team of experts was divided so that two experts shared the annotation workload of each assigned class. Each image was assigned a label by one annotator only. Due to time limitation, we were unable to test for consistency between annotators by allowing multiple annotators to label the same subset of images, although that would be a useful additional step should resources allow. The workshop included thirty minute to one hour discussion sessions at the start and the end of each day and was structured so that one-hour annotating sessions were preceded by thirty minute scientific presentations by selected workshop participants. This format is not prescriptive, but was successful in keeping the community engaged in our case.

How well did we do?

Successes and limitations of workshop design

The workshop setting allowed us to design a rigorous set of taxonomic and semantic classes, collect plankton images for training and build a first-cut classifier in a matter of days. Focused discussions among the experts provided valuable insight into what organisms can be confidently identified from the images collected by the SPC. In addition to image annotation sessions, the workshop facilitated scientific exchange and introduction to machine learning, resulting in seven hours allocated for image annotation alone. Expert annotators classified a total of 18,303 images into 32 taxonomic and three noise categories (listed in Fig. 3a).

Annotation of 18,303 images by 17 participants over the course of the workshop may not seem like a lot, in comparison to the rate of image acquisition reported elsewhere (Lombard et al. 2019). In our annotation approach, once the database was populated with the satisfactory number of example images for a particular classification category (here the goal was 500 images per class), the expert would then proceed to annotate images for another category to ensure that sufficient number of images is collected for each class. Annotation of images for highly abundant organisms or particles was conducted very swiftly, however, the annotation rate dropped significantly when looking for images of rare organisms. The merit of our approach is that it ensures less abundant organisms are equally well represented in the image database.

During the workshop, we encountered a few issues to consider in future annotation efforts: (1) differences in the sampling techniques between the SPC and the weekly pier monitoring data; (2) annotator recency bias; and (3) the annotators' ability to set search parameters to obtain unlabeled images and create new class labels in the data browser.

Selection of dates for image annotation of each target taxon was to a large extent informed by the ecological data provided by HABMAP. Selected dates were biased toward periods of elevated abundance of a particular taxon as reported by HABMAP. During the proceedings, it became apparent that, for some organisms, high abundances indicated in the weekly HABMAP samples were not reflected in the image data collected by the SPC (and vice versa). These discrepancies may have been due to different sampling designs, namely, the depths of water collection (HABMAP samples are collected from the surface, whereas SPC is deployed at about 3 m). The discrepancies were most pronounced for rare organisms, making it challenging to collect enough training images for less abundant taxa. Thus, for the limited number of taxa for which our image collection procedures did not yield a sufficient number of images, we decided to focus our efforts on the limited periods when the SPC captured many images of a particular taxon, hence sacrificing the temporal distribution of training images to guarantee a sufficient number of training examples. Collection of images of rare species remained a challenge, with a handful of taxonomic classes having to be excluded from automated classification due to an insufficient number of training images.

Another pervasive issue to consider is recency bias: annotators would subconsciously assume a higher likelihood of a new image belonging to a taxonomic class that had a high relative abundance in a discrete time period (Culverhouse et al. 2014). Annotators assigned images to a specific class even when there was an insufficient level of structural detail to enable confident taxonomic identification. Inspection of training images collected during the workshop revealed a substantial number of images whose assignment to a particular taxonomic class was questionable.

Finally, annotators had full access to search parameters in the web interface and the flexibility to create new labels. Even though the classification categories were preset, the ability to create new labels aimed to highlight new or interesting organisms that were not accounted for in our annotation scheme, and somehow resulted in multiple labels for predetermined classification categories. This source of confusion could have easily been avoided, either by limiting user privileges or fixing the search parameters on the user interface. Ideally, the interface should be designed so that annotators can focus purely on image classification and avoid distractions associated with interface setup.

The image dataset annotated during the workshop was manually quality controlled afterwards by a single person with taxonomic, database, and machine learning expertise. The postprocessing included: (1) clearing out inconsistencies in label assignments; (2) removal of misclassified images; (3) further differentiation of taxonomic classes that combined multiple taxa; (4) introduction of semantic descriptors to further separate highly morphologically diverse classes (e.g., *T. furca* was often imaged either in a form of a single cell or as a pair of cells and so the class was further subdivided using semantic descriptors "single" and "pair" to distinguish the two morphologies); and (5) merging classes of organisms that were very similar morphologically, and therefore not distinguishable if imaged at suboptimal orientation (e.g., *Rhizosolenia* and *Proboscia* spp. that are morphologically similar solitary diatoms). The image database and the classifier continued to be developed after the workshop to improve the classifier performance, where the latest version will be made available at https://doi.org/10.6075/J00865GT.

What would we change now?

We learned several key lessons during our workshop that should improve annotation efficiency in future efforts.

It is important that the principal and only responsibility of taxonomy experts is image annotation. Programmatic creation of new labels and setting of search parameters (e.g., date, time, size range, etc.) should be limited to the annotation software administrator to ensure consistency. This can be done by compiling all images within the preselected time intervals and size range prior to the annotation sessions, equally

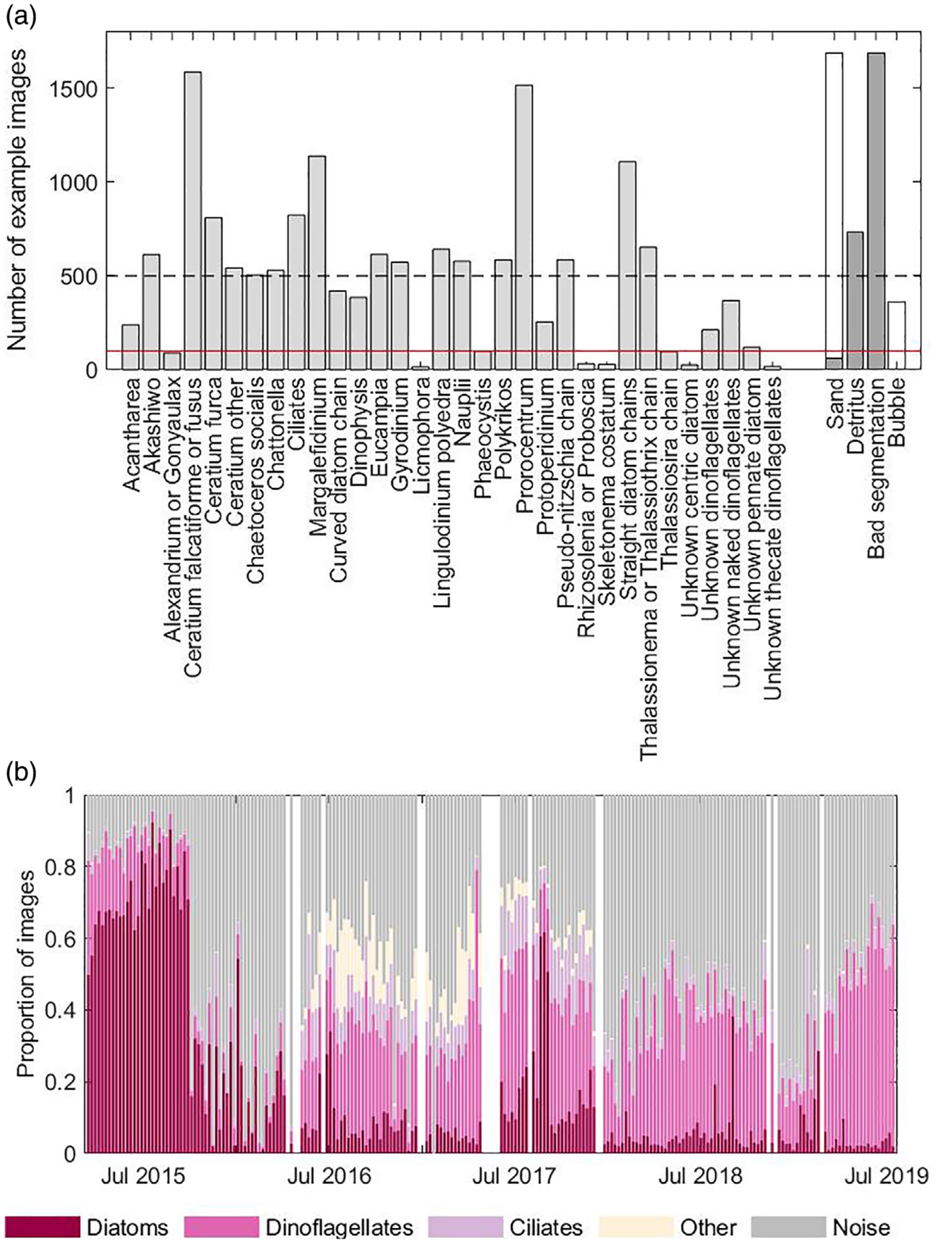

FIG. 3. (**a**) The number of images collected per each category for training of the automated classifier. Taxonomy-based categories are in light gray, and non-biological (i.e., "noise") categories are in dark gray. The black, dashed line marks the goal for image annotation. The red line marks the minimum of images required for a category to be included in the automated classification. Note that for categories "Sand" and "Bubble," additional images were collected prior to the workshop, indicated by white bars, that were used for training of the classifier. (**b**) The output of the classifier applied to image data collected every Monday, from March 2015 to July 2019, illustrating the proportion of all images classified cumulatively into diatom, dinoflagellate, ciliate, other and noise categories. Category "Other" here includes "Chattonella," "Nauplii," and "Acantharea" categories.

dividing the images among annotators and providing each expert with their set of images for annotation. Preselecting images and limiting the annotator's interaction with the software would be more time efficient.

Preselecting data would also allow organizers to mix images from different dates, helping to address recency bias and increasing the temporal distribution of the training set. Rather than focusing on one class at a time, annotators would then be identifying a multitude of taxa from a greater variety of time intervals, forcing them to pay attention to important morphological features. Moreover, this approach would ensure an equal distribution of labor among annotators and limit fatigue related to hunting for images of rare species.

Previous research has demonstrated that human annotators suffer from biases and are often not consistent with their own previous effort (Culverhouse et al. 2014), highlighting the need for some sort of quality control of the image database. We attempted to account for annotator biases by having a single expert taxonomist quality control all annotations. Ideally, to evaluate these biases, the level of agreement between annotators should be assessed by selecting a subset of images that will be examined by all experts individually (Schoening et al. 2016). Comparing the outcome of such simultaneous annotations would provide confidence in the consistency of the labeled image data set and highlight any problematic taxa if there was high discrepancy among the experts with varying levels of experience.

There are numerous options for image annotation software and services available for data of all kinds (Gomes-Pereira et al. 2016; Irisson et al. 2022). Some are freely available on GitHub. Others, like MATLAB's Image Labeler GUI or LabelBox, are professionally supported by large, for-profit companies. There are several ocean-specific annotation tools like Ecotaxa (https://ecotaxa.obs-vlfr.fr/), BIIGLE (https://biigle.de/), VARS (https://www.mbari.org/technology/video-annotation-and-reference-system-vars/), or IFCB Annotate (https://ifcb-annotate.whoi.edu/), to name just a few. All of the above software are powerful tools for data annotation; however, selection of a suitable software should be made based on the annotation design to best fit the project's needs. Among other considerations, users should pay attention to, for example, the output format of the labels, whether the package is actively supported, and options for integrated data management.

Collection of images of rare species remains a challenge. One method to address this issue might be to monitor the image database until a sufficient number of images for training is collected. Alternatively, application of unsupervised clustering algorithms, such as Morphocluster (Schröder et al. 2020), can be incredibly helpful for limiting the number of images for inspection. One could also use the classifier as a prefilter, following the principles of active learning (Settles 2009): iteratively running a trained classifier, inspecting the output in search of correctly identified images of rare taxa, and retraining the classifier with a larger training set. However, it is important to note that this approach may ultimately reinforce what the classifier has already learned about a given class. Care must be taken to ensure the computer learns what it is getting wrong by properly sorting false positives: images that were incorrectly classifier should be manually reclassified and included in the training of the next classifier iteration.

But the outcomes!

Plankton time series

Annotation of plankton images during the workshop enabled training and application of the first-cut classifier to the remaining images collected by the SPC. During the workshop, only images collected on Mondays were classified in order to decrease the computational time while the classifier continues to be further developed for improved performance. Here, we illustrate the data product that resulted from the efforts invested by the scientists during the workshop: the classifier output (Fig. 3), where the taxonomic categories were grouped into functional groups (diatoms and dinoflagellates) as the output requires further, class-specific validation. The image dataset and the classifier underwent substantial development after the workshop and some of the final ecological time series have been validated and published (Kenitz et al. 2023).

It was worth it!

Novel aquatic imaging technologies have enormous promise for speeding the acquisition of data on plankton community composition and dynamics, potentially allowing scientists to monitor ecosystems at high spatial and temporal resolution and ask exciting new scientific questions (Irisson et al. 2022). Converting the overwhelming amounts of image data into ecologically meaningful information requires machine learning and computer vision algorithms. Collecting the large volume of image data required for training of these algorithms is time consuming, challenging, and expensive for complex planktonic communities, where intra-species morphological diversity and multiple life cycle stages may be present. The morphological differences between species or classes can be small and only distinguishable by experts.

A workshop setting has proven to be a productive and efficient approach for soliciting scientific expertise and speeding collection of high-quality, labeled data required for training of automated classifiers. Focused discussions facilitated quick decisions and consensus on which organisms are most relevant for regional monitoring efforts and the feasibility of their taxonomic identification from images. This pivotal part of automated classification workflow is often the most challenging and time-consuming, irrespective of the imaging technology and image annotating software used.

Planktonic communities are exceptionally dynamic. Community structure and even cell morphology can vary considerably in both time and space. Automated classifiers often assume static relationships between data used for their training and the variability observed in situ, which can introduce significant errors to the final data output. It is important to carefully monitor the performance of automated classifiers, highlighting the value of expert involvement in the continuous validation of the classifier output. We believe the most effective way of ensuring consistent, high-accuracy output is maintaining human expert participation in all stages of the automated classification process.

Acknowledgments

Sustained phytoplankton monitoring efforts at Scripps Pier are conducted by the Southern California Coastal Ocean Observing System (Award # NA16NOS0120022) as a part of the California Harmful Algal Bloom Monitoring and Alert Program (HABMAP), monitoring water quality along the central and southern California coast. The workshop was funded by the National Science Foundation BIGDATA Initiative award (NSF IIS 15-46351) to JSJ and PJSF. KMK and ADB were supported by the Simons Foundation. We thank the participants of the taxonomy workshop for their participation and engagement. The Olympic Region Harmful Algal Bloom (ORHAB) partnership has been funded since 2006 by a surcharge to Washington State shellfish and seaweed licenses as authorized by RCW 77.21.555. We thank ORHAB for providing support for Anthony Odell.

Author contributions

KMK and ECO organized the Taxonomy Workshop at Scripps Institution of Oceanography. KMK, ECO and ADB prepared the manuscript. All authors participated in the workshop and made significant contributions to the content of this manuscript.

References

Axler, K. E., S. Sponaugle, C. Briseño-Avena, F. Hernandez, S. J. Warner, B. Dzwonkowski, S. L. Dykstra, and R. K. Cowen. 2020. Fine-scale larval fish distributions and predator−prey dynamics in a coastal river-dominated ecosystem. Mar. Ecol. Prog. Ser. **650**: 37–61. doi: 10.3354/meps13397.

Culverhouse, P. F., N. Macleod, R. Williams, M. C. Benfield, R. M. Lopes, and M. Picheral. 2014. An empirical assessment of the consistency of taxonomic identifications. Mar. Biol. Res. **10**: 73–84. doi:10.1080/17451000.2013.810762.

Gomes-Pereira, J. N., and others. 2016. Current and future trends in marine image annotation software. Prog. Oceanogr. **149**: 106–120. doi: 10.1016/j.pocean.2016.07.005.

González, P., E. Álvarez, J. Díez, Á. López-Urrutia, and J. J. del Coz. 2017. Validation methods for plankton image classification systems. Limnol. Oceanogr.: Methods **15**: 221–237. doi:10.1002/lom3.10151.

Irisson, J.-O., S.-D. Ayata, D. J. Lindsay, L. Karp-Boss, and L. Stemmann. 2022. Machine learning for the study of plankton and marine snow from images. Ann. Rev. Mar. Sci. **14**: 277–301. doi:10.1146/annurev-marine-041921-013023.

Kenitz, K. M., E. C. Orenstein, P. L. D. Roberts, P. J. S. Franks, J. S. Jaffe, M. L. Carter, and A. D. Barton. 2020. Environmental drivers of population variability in colony-forming marine diatoms. Limnol. Oceanogr. **65**: 2515–2528. doi:10.1002/lno.11468.

Kenitz, K. M., and others. 2023. Environmental and ecological drivers of harmful algal blooms revealed by automated underwater microscopy. Limnol. Oceanogr. **68**: 598–615. doi: 10.1002/lno.12297.

LeCun, Y., Y. Bengio, and G. Hinton. 2015. Deep learning. Nature **521**: 436–444. doi: 10.1038/nature14539.

Li, Y., and N. Vasconcelos. 2020. Background data resampling for outlier-aware classification. Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. 13218–13227. doi:10.1109/CVPR42600.2020.01323.

Lombard, F., and others. 2019. Globally consistent quantitative observations of planktonic ecosystems. Front. Mar. Sci. **6**: 196. doi: 10.3389/fmars.2019.00196.

Orenstein, E. C., K. M. Kenitz, P. L. Roberts, P. J. S. Franks, J. S. Jaffe, and A. D. Barton. 2020*a*. Semi- and fully supervised quantification techniques to improve population estimates from machine classifiers. Limnol. Oceanogr.: Methods **18**: 739–753. doi:10.1002/lom3.10399.

Orenstein, E. C., D. Ratelle, C. Briseño-Avena, M. L. Carter, P. J. S. Franks, J. S. Jaffe, and P. L. D. Roberts. 2020*b*. The Scripps plankton camera system: a framework and platform for in situ microscopy. Limnol. Oceanogr.: Methods **18**: 681–695. doi:10.1002/lom3.10394.

Schoening, T., J. Osterloff, and T. W. Nattkemper. 2016. RecoMIA-recommendations for marine image annotation: lessons learned and future directions. Front. Mar. Sci. **3**: 59. doi: 10.3389/fmars.2016.00059.

Schröder, S. M., R. Kiko, and R. Koch. 2020. MorphoCluster: Efficient annotation of plankton images by clustering. Sensors **20**: 3060. doi:10.3390/s20113060.

Settles, B. 2009. Computer sciences active learning literature survey. Compututer Science Technical Report 1648.

Sosik, H. M., and R. J. Olson. 2007. Automated taxonomic classification of phytoplankton sampled with imaging-in-flow cytometry. Limnol. Oceanogr.: Methods **5**: 204–216. doi: 10.4319/lom.2007.5.204.

Kasia M. Kenitz, Scripps Institution of Oceanography, University of California San Diego, La Jolla, CA; kkenitz@ucsd.edu

Eric C. Orenstein, Scripps Institution of Oceanography, University of California San Diego, La Jolla, CA

Clarissa R. Anderson, Scripps Institution of Oceanography, University of California San Diego, La Jolla, CA

Alexander J. Barth, Department of Biological Sciences, California Polytechnic State University, San Luis Obispo, CA; Biological Sciences, University of South Carolina, Columbia, SC

Christian Briseño-Avena, Department of Environmental and Ocean Sciences, University of San Diego, San Diego, CA

David A. Caron, Department of Biological Sciences, University of Southern California, Los Angeles, CA

Melissa L. Carter, Scripps Institution of Oceanography, University of California San Diego, La Jolla, CA

Emily Eggleston, Department of Biological Sciences, University of Southern California, Los Angeles, CA

Peter J. S. Franks, Scripps Institution of Oceanography, University of California San Diego, La Jolla, CA

James T. Fumo, Scripps Institution of Oceanography, University of California San Diego, La Jolla, CA; University of Hawai'i at Manoa, Honolulu, HI

Jules S. Jaffe, Scripps Institution of Oceanography, University of California San Diego, La Jolla, CA

Kelsey A. McBeain, University of Hawai'i at Manoa, Honolulu, HI

Anthony Odell, University of Washington's Olympic Natural Resources Center, Forks, WA

Kristi Seech, Scripps Institution of Oceanography, University of California San Diego, La Jolla, CA

Rebecca Shipe, Institute of the Environment and Sustainability, University of California Los Angeles, Los Angeles, CA

Jayme Smith, Southern California Coastal Water Research Project Authority, Costa Mesa, CA

Darcy A. A. Taniguchi, Biology Department, California State University San Marcos, San Marcos, CA

Elizabeth L. Venrick, Scripps Institution of Oceanography, University of California San Diego, La Jolla, CA

Andrew D. Barton, Scripps Institution of Oceanography, University of California San Diego, La Jolla, CA; Department of Ecology, Behavior and Evolution, University of California San Diego, La Jolla, CA