Abstract—Computational archival science (CAS) provides new pathways for research. Biologists, for example, can perform scientific studies by applying AI/ML to digital biological specimen collections and explore questions that were not possible in the analog world. One such approach is the application of computational methods for specimen outlining to assist with specimen identification, morphometry, and other scientific questions. The challenge is to determine how to computationally generate and represent a specimen’s outline. The research presented in this paper addresses this challenge, through the deployment of elliptical Fourier descriptors (EFDs). The paper describes the image processing pipeline for extracting fish outlines, a key morphological feature, and representing the outlines using EFDs. In addition, our research presents the application of machine learning classification on the EFDs. The resulting dataset is well suited for a variety of machine learning-based downstream analyses, including classification by genus and species. Overall, the classification tests produced a 96.3% accuracy, demonstrating the distinguishing nature of the EFDs, and by proxy, the fish outlines as a whole. Broadly, these results indicate the effectiveness of archival specimen usage in machine learning applications, and demonstrate specimen outlining via Fourier descriptors as a computational archival science approach.

I. INTRODUCTION

Records are created during everyday transactions. They are frequently generated through machine-driven processes in our increasingly digital, networked world, although there are still many record examples generated through some combination of human and machine-driven operations. One may consider an invoice from a doctor’s office visit, a record of an online Amazon purchase, or documentation confirming completion of a certification credential. Records are generated in multiple formats. Images, specifically specimen images, are one of the most significant record types in biological research. A type-specimen serves to coin a species name, and specimen images allow biologists to study a new species, explore species variations, and make other scientific discoveries. Like many records, specimen images housed in archives and digital repositories serve as archival records and can aid in scientific discovery.

As Artificial Intelligence/Machine Learning (AI/ML) methods are increasingly being explored across nearly every research domain, it is not surprising that scientists have become curious about how such methods can leverage digital, archival specimen repositories for scientific research [1]. Computational archival science (CAS) [2], [3] provides a pathway for exploiting archival specimens, applying computational thinking, and supporting research that spans the humanities [4] to the biological sciences [5]–[7] and other domains.

Applying computational techniques to these archival records to automatically extract important metadata is a key task for preparing archives for future AI/ML analysis [8]. Over the last couple of decades, there has been significant effort to digitize museum specimens [9], [10], but many of the resulting images lack the metadata needed for downstream AI/ML analysis. Many techniques therefore need to be developed and applied to scientific images in order to make the related specimen archive records “AI-ready”. We demonstrate one such computational technique, the automated outlining of objects in museum specimen images. The outline is then transformed into a mathematical representation that is suitable for downstream computational processing. Having these types of tools identified, developed, and deployed will advance the exploitation of the vast stores of information available in museum archives, leading to future scientific discoveries based on this previously inaccessible information.

In our work, we employ a set of image processing techniques to segment and outline fish found in museum archive images. The outlines are converted into elliptical Fourier descriptors (EFDs) [11], which transform the 2D shape into a 1D vector that may then be analyzed with machine learning techniques. We highlight the feasibility and value of the approach, and the effectiveness of EFDs specifically, in a classification task that identifies the genus of the fish with a...
high accuracy rate simply from their 2D outline shape. These results demonstrate that the computationally derived archival data is useful for advancing biological research.

The sections that follow cover background work on metadata and elliptical Fourier descriptors, present our research aims, detail our computational methods, and report our results and conclusions.

II. RELATED WORK

A. Metadata for Scientific Specimens

Metadata has always been an essential component in the capture, organization, and preservation of scientific specimens. Metadata labels provide crucial specimen data, the specimen’s genus and species, the date the specimen was produced, who collected and identified the specimen, and the geo-location documenting where the specimen was obtained. Metadata has been essential in supporting specimen access, preservation, and use when transforming analog herbaria and museum collections to digital forms. The Darwin Core metadata standard [12], [13] was developed specifically for recording specimen label information in digital form. Additional metadata standards, such as Preservation Metadata Implementation Strategies (PREMIS) [14] for preservation metadata and the Exchangeable Image File Format (EXIF) for image metadata, have also played a key role in recording preservation characteristics of the original analog specimen and technical details associated with the digitization process. Another key standard, PROV [15], has been important for tracking historical information about specimen collections.

Although the menu of metadata standards is extensive, existing standards fail to sufficiently support rigorous scientific research. The result is digital repositories containing hundreds of thousands of digital scientific specimens with inadequate metadata. The images take on an archival existence, in that they are maintained due to their enduring and evidential value [16] – a mainstay of archival practice. However, there is a pressing need for more granular metadata to aid scientific research [17]. One key need expressed by scientists is metadata specific to the 2D shape of a specimen. Such metadata is central to identifying species class and if it could be computationally generated, it would aid in scientific research. Elliptical Fourier descriptors can address this need, and is the focus of the CAS research presented in this paper.

B. Elliptical Fourier Descriptors for Scientific Studies

Elliptical Fourier descriptors (EFDs) have been used in a number of biological scientific studies. Rohlf and Archie [18] explored the ability of various Fourier methods to capture the shape of different mosquito wings. They found the use of EFDs to be especially promising for representing the pattern of similarities and differences in wing shapes. Diaz et al. [19] used EFDs to characterize cell and nuclear shape and show that different types of cells can be differentiated based on EFD harmonics. Iwata et al. [20] quantitatively evaluated EFDs’ ability to capture variations in the root shape of the Japanese radish, due to genotypes, soil types and growth stages. Varietal influence was shown to be strong, indicating that the represented traits are highly heritable and may be effectively selected.

Soares et al. [21] carried out a pilot study of leaf outline morphotypes on populations of five species of Hetempsis located in the Brazilian Amazon. The aim of the study was to investigate quantitative methods for discriminating species within a local area based on vegetative morphology, which could be valuable for conservation of natural populations. Mebatsion, Paliwal and Jayas [22] employed EFDs to study the variation in the shape of cereal grains, namely barley, oat, rye, and wheat. EFDs were effective in capturing shape differences between different growing regions. Demir et al. [23] showed that EFDs could be used to distinguish between types of almonds. Automatically identifying almond type is important for automated processing of the nuts.

Triatoma dimidiata is an insect that spreads Chagas disease in various countries in the Americas. Methods for identifying them using morphometric analyses with landmarks have not yet been fully resolutive. The work of Cruz et al. [24] validated the use of EFDs for the identification of three lineages of this species complex. In an interesting non-biology application, Dhingra et al. [25] used EFDs to quantify the shapes of Titan’s lakes. They concluded that the method shows promise for testing hypotheses for Titan lake formation, because they found that Titan’s lakes could be quantitatively placed in four distinct shape-based groups.

Our work goes beyond previous work in that we have developed an improved method for generating the initial 2D outline of the studied specimen, and we employ modern AI-based methods for detecting and segmenting specific objects before applying EFD analysis.

III. RESEARCH AIMS

The overall goal of this work is to further increase the scope of our metadata extraction capabilities for museum specimen images, as a computational archival science approach. Our technique brings us closer to this goal, and demonstrates how we can gather more comprehensive metadata for each digital specimen image and expand their utility for scientific inquiry. This work has three specific research aims:

- Develop an improved image processing technique for extracting outlines from museum specimen images,
- Explore a distinguishing specimen outline representation based on elliptical Fourier descriptors, and methods for maximizing its distinguishability,
- Demonstrate the effectiveness of our computational archival science approach with an example downstream analysis which is able to predict the genus of fish specimens with high accuracy.

The following section details our methods for 2D shape outlining, shape representation, and machine learning analysis.
IV. METHODS

The high-level steps of our approach are:

- Segment the fish specimen image into its separate elements: fish, ruler, and information card.
- Compute the outline of the fish.
- Convert the pixels of the outline into a numerical shape description, specifically elliptical Fourier descriptors (EFD). This conversion transforms the 2D outline into a 1D form that may be used for machine learning analysis.
- Perform a classification task on the EFD-based feature vector to demonstrate its usefulness in downstream AI applications.

A. Segmentation of Fish Specimen Images

As reported in earlier work [5], [6], the first step in an automated metadata generation process for museum specimen images is finding the specimens (and other relevant objects) within the collection images. Object detection is now a broadly active field of study [26], and has resulted in a number of well-tested, purpose-built architectures. We elected to use Facebook AI Research’s (FAIR) detectron tool [27], and specifically its implementation of the Mask R-CNN architecture [28], for object detection in our work, given its many flexible and robust capabilities. Figure 1 presents typical output from detectron, which highlights the computed bounding boxes around the detected fish, ruler, and information card in the image.

B. Fish Outline Extraction Algorithm

Given the bounding box of the fish in the specimen image, the following are the detailed computational steps of the fish outline extraction algorithm. These steps are pictorially summarized in Figure 2 and can be organized into four major procedures: Identify Fish Pixels, Segmentation Clean Up, Standardize Orientation and Size of Fish, and Create Standardized Outline from Pixels.

Identify Fish Pixels
- Expand the fish’s bounding box by 5% to ensure that the whole fish is circumscribed and extract the sub-image of the fish.
- Convert the fish image from RGB to HSV.
- Perform Otsu thresholding [29] on the Saturation channel of the fish image.
- Define the background as those pixels with values below the Otsu threshold.
- Calculate the mean and standard deviation of the dark, background pixels.
- Reset threshold to be the mean + c \times \text{ standard deviation}. We use a c value of 0.5 to generate our results. Shifting the threshold closer to the mean of the dark pixels has shown to improve the accuracy of the fish outline.
- Perform a binary segmentation of the fish image using the adjusted threshold value to produce a mask of the fish. White pixels are foreground (fish) and black pixels are background.

Segmentation Clean Up
- In some cases other elements of the fish specimen image, e.g. the information card, may overlap with the fish. This corrupts the downstream calculations. We then detect if the bounding boxes of the other elements overlap the bounding box of the fish. The pixels of the overlapping non-fish bounding boxes are set to black.
- Find every connected component in the image and black out all but the two largest components by area. These components will usually be the fish and the background. Any smaller components are unwanted image artifacts.
- Perform a morphological closing to smooth the boundary of the fish. This removes high-frequency noise along the boundary.

Standardize Orientation and Size of Fish
- Compute the fish’s primary axis as the first principal component derived from Principal Component Analysis applied to the fish pixels.
- Rotate the fish pixels about their centroid, so the fish’s primary axis is parallel with the horizontal axis.
- If the metadata from [5] indicates that the fish faces right, then mirror the image about the vertical axis. The follow-on analysis assumes that all fish face to the left.
- Using the pixels/cm metadata item from [5], scale the fish images, so they have a consistent physical size. This scaling step ensures that the EFDs not just encode shape, but also include size information. We have found that having the fish representations be approximately consistent in physical units improves classification results downstream.

Create Standardized Outline from Pixels
- Outline the resulting fish component using the Suzuki and Abe edge detection algorithm [30].
- The outline is structured as a counter-clockwise-ordered list of pixels on the fish component boundary, with the first element in the list being the left-most pixel. This constraint produces a consistent outline representation for
all fish.
- Subtract the fish’s centroid from each outline coordinate. This ensures that the fish outline encircles the origin, a property needed for the EFD transformation.
- Remove any duplicate points in the outline pixel list.

C. Elliptical Fourier Descriptors Generation

The next stage of the computational process converts the list of pixels that describe the outline of the fish specimen into a mathematical form that is more amenable for downstream machine learning analysis. We use elliptical Fourier descriptors (EFDs) [11] to represent the shape of the outline. Converting 2D outlines into EFDs transforms a list of 2D locations into a list of (scalar) elliptical Fourier coefficients. This list can be processed into a feature vector that captures the shape characteristics of the outline.

The mathematical definition of a 2D outline in EFD form is

\[
x(t) = \sum_{n=1}^{N} \left[ A_n \cos \left( \frac{2\pi nt}{T} \right) + B_n \sin \left( \frac{2\pi nt}{T} \right) \right]
\]

(1)

\[
y(t) = \sum_{n=1}^{N} \left[ C_n \cos \left( \frac{2\pi nt}{T} \right) + D_n \sin \left( \frac{2\pi nt}{T} \right) \right]
\]

(2)

where \( n \) = the harmonic number, \( N \) = the maximum harmonic number, \( t \) = distance along the outline, and \( T \) = total length of the outline. These equations define a univariate parametric closed curve. Instead of utilizing polynomials, which are more
typically used [31]. EFDs employ a Fourier series, a finite sum of sines and cosines at different frequencies, to compute an \((x, y)\) for every \(t\) value along the length of the outline. \(A_n, B_n, C_n, D_n\) are the EFD coefficients associated with each harmonic \(n\). EFDs produce an approximation to the original outline, with the fidelity of the reconstructed outline improving as \(N\) increases.

Kuhl and Giardina [11] described methods for computing \(A_n, B_n, C_n, D_n\) for a given outline and a specific value of \(N\). We employ the implementation of their method available in the PyEFD library [32]. Computing \(A_n, B_n, C_n, D_n\) from the extracted fish outlines transforms the 2D sequence of points into a 1D vector of EFD coefficients. \(N\) is adjusted for each fish image such that the associated reconstructed outline falls within 1mm, measured with the symmetric Hausdorff distance, of the original outline. This ensures a consistent error tolerance across all fish specimen EFDs.

Additional transformations are applied to the raw EFD coefficients vector, to make the embedded information more tractable to analysis. As recommended by Kuhl and Giardina [11], the EFD coefficients are normalized, such that \(A_0 = 1, B_0 = 0\) and \(C_0 = 0\). This makes the first harmonic define an ellipsoid with an \(x\) axis of length 1. Since EFD normalization makes the first three coefficients the same for all EFDs, these coefficients may be removed from the vector.

Initially, the maximum \(N\) was set to 78, producing up to 312 EFD coefficients for some cases. We found that these higher harmonics frequently included noise that lowered our subsequent classification accuracy. We performed a parameter study and found that a maximum value of \(N\) equal to 31 produced the highest accuracy results. If a specimen’s \(N\) was greater than 31, the higher harmonics are removed. If a specimen’s \(N\) is lower than this value, then the coefficient vector is padded with zeros, so that all EFD coefficient vectors have the same length. Next, Z-score normalization is performed, which shifts and scales the EFD coefficients to have a mean of 0 and a standard deviation of 1. Finally, linear discriminant analysis (LDA) is applied to distill down the normalized EFD coefficients to a 6D feature vector.

**D. Genus Classification**

In order to demonstrate the expressiveness and usefulness of the derived archival records, the outlining techniques were applied to a set of images of fish from different genera. The resulting feature vectors were used in a classification task to predict the genus of a fish from its 2D outline. A number of classification methods were applied to our data, namely Support Vector Machines (SVM) [33], Multi-layer Perceptrons [34], XGBoost [35] and K-Nearest-Neighbor [36]. While all methods produced acceptable results, we found that SVM produced predictions with the highest accuracy.

**V. RESULTS AND DISCUSSION**

Our outlining method was applied to 1,071 fish images from the Illinois Natural History Survey (INHS) [38]. The dataset consists of 7 genera (Notropis, Phenacobius, Gambusia, Cyprinus, Esox, Lepomis, Noturus), with each genus containing 153 specimens. 5-fold cross-validation was used with SVM to predict the genus of all specimens in the dataset. The overall prediction accuracy for the task was 96.3 ± 1.5%.

Although we made important progress, it is important to point out that our EFD-based feature representation is not easily interpretable. Interpretability is a valuable quality of a shape representation because it can help scientists identify the specific parts of a fish’s outline that distinguish its genus or species. Our current outline representation is highly distinguishable, but more research is needed to achieve a higher level of interpretability. For example, perturbing values in an EFD encoding affects the reconstructed outline unintuitively. Furthermore, the resulting shape, while useful, will likely not fully resemble a fish. It is not straightforward to, for example, design a vector that we can add to a fish’s coefficients to double the fish’s dorsal fin height. Such a vector would differ for every fish.

Despite these limitations, we recognize that extracting 2D shape outlines, represented with EFDs, from specimen images provides useful quantitative information. EFDs’ ability to capture identifying data is clearly evident in a UMAP visualization [37] of the fish specimen feature vectors. UMAP (Uniform Manifold Approximation and Projection) is a manifold learning technique for dimension reduction. UMAP is constructed from a theoretical framework based on Riemannian geometry and algebraic topology. The method is able to map points in a high dimensional space down into 2D, while maintaining the global topological structure of the original data. Figure 3 presents a 2D map of the 6D feature vectors for the seven genera of fish in our dataset. Each 2D data point, which represents a fish specimen, is color-coded by genus. It can be seen that the feature vectors based on EFDs form clusters by genus within the feature space, with just a few exceptions.
VI. CONCLUSION

We have developed a computational method for automatically outlining specimens in museum images. The 2D outlines are then transformed into a 1D feature vector using elliptical Fourier descriptors (EFDs). 2D outlines and their mathematical representation are one of many types of metadata that can be extracted from museum archives for use in scientific studies. We have demonstrated the ability of computational outlining to make the related specimen archive record "AI-ready". The high accuracy of our genus predictions highlights the expressiveness and usefulness of the EFDs derived from the 2D geometric outlines of the specimens. Overall, the research presented here demonstrates the promise of computational techniques to support researchers in leveraging and exploiting archival collections for scientific discoveries.

ACKNOWLEDGMENT

We would like to thank Yasin Bakiş, Xiaojun Wang and Henry Bart Jr. of the Tulane University Biodiversity Research Institute for their support and fish biology expertise.

REFERENCES