DREXEL UNIVERSITY
College of
Computing &
Informatics

DREXEL UNIVERSITY
Metadata
Research Center
College of Computing & Informatics

Specimen Outlining:

A Computational Archival Science Approach

David E. Breen, Andrew Senin, Ajani Levere, Joel Pepper, Jane Greenberg

Metadata Research Center, Drexel University

December 15-18, 2023

Outline

01

Introduction

- Background
- Research Aims
- Workflow Overview

02

Methodology

- Fish Segmentation
- Outline Extraction
- EFD Generation, Feature Engineering
- Genus Classification

03

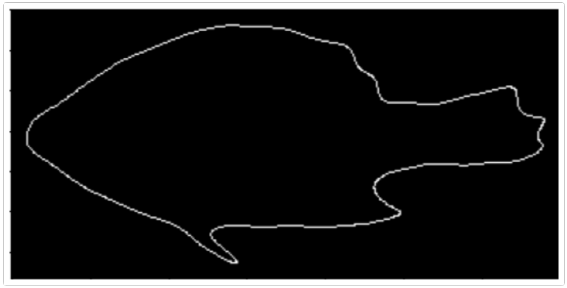
Results

04

Conclusion

Background

- Proliferation of digital record creation increases significance and feasibility of exploiting record types, particularly specimen images
- Specimen images possess unique metadata, study of such attributes further scientific discovery
- Modern AI/ML methods can gain new knowledge from large archives' metadata [1], but such archives are not readily usable, i.e., "AI-ready"



Background (cont.)

We present a computational technique demonstrating the automatic outlining of museum specimen images

Research Aims

As part of ongoing efforts to develop methods for extracting AI-ready metadata profiles of museum specimens, we

- Present an improved image processing technique for extracting specimen outlines, a useful feature for biological research
- Explore distinguishing specimen outline representation via Elliptical Fourier Descriptors (EFDs)
- Demonstrate the effectiveness of the approach via downstream analyses (genus classification and UMAP)

Workflow Overview

- 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)
    - 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

Methods: Fish Segmentation

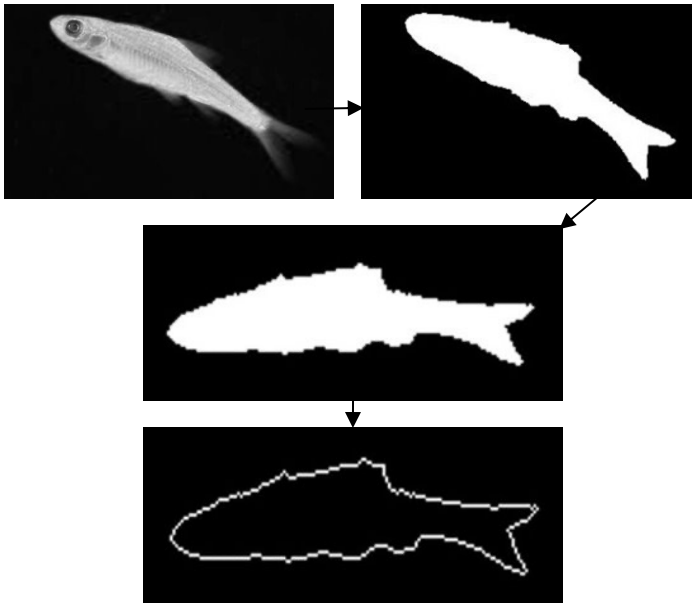
We trained Facebook AI Research's Detectron tool [6] on 100 hand-labeled images from the Illinois Natural History Survey (INHS). Then we used it to extract from each image:

- Specimen region of interest
- Image spatial resolution (pixels/cm) [4]
- Specimen genus and species [3]

Methods: Outline Extraction

Continuing with the specimen ROI, we

- Convert to HSV color space
- Discard H and V channels, yielding a "saturation image"
- Binarize image on an adjusted Otsu threshold [5]
- Normalize scale and orientation
- Obtain ground truth outline from normalized shape



Methods: EFD Generation

Elliptical Fourier Descriptors (EFDs) [2] transform a list of 2D locations into a list of (scalar) elliptical Fourier coefficients

- Utilize Fourier series
- Consist of some number of harmonics
- Approximate the ground truth outline with increasing accuracy for every additional harmonic

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

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

$$(x_1, y_1), (x_2, y_2), ..., (x_n, y_n) \longrightarrow a_1, b_1, c_1, d_1, a_2, b_2, c_2, d_2, ..., a_n, b_n, c_n, d_n$$

Methods: EFD Generation (cont.)

5mm tolerance, 1 harmonic

2.5mm tolerance, 3 harmonics

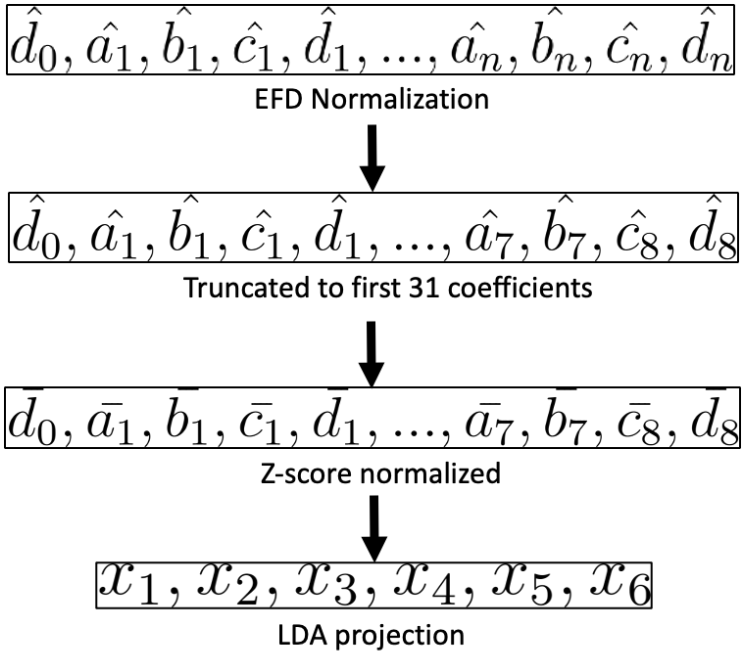
2mm tolerance, 4 harmonics

1.5mm tolerance, 6 harmonics

1mm tolerance, 10 harmonics
(used in this study)

0.5mm tolerance, 15 harmonics

Methods: Feature Engineering



Various transformations were applied to EFD vector data

- EFD Normalization – Removal of 3 coefficients from vector [2]
- N = 31 – Reducing noise, increasing subsequent classification results
- Z-Score Normalization – EFD coefficients to have mean of 0 and standard deviation of 1
- Linear Discriminant Analysis (LDA) – distills normalized coefficients to 6D feature vector

Methods: Genus Classification

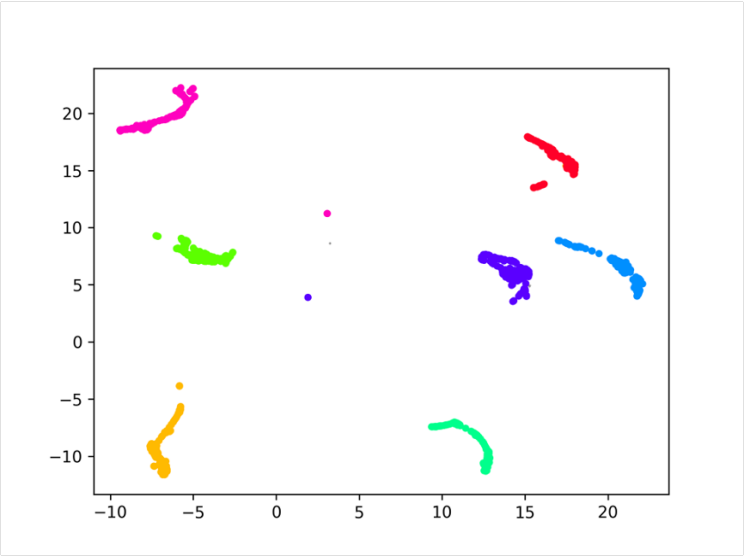
- To demonstrate expressiveness of archival records, outlining technique was applied on images of fish from various genera
- Allowed for classification via genus given a 2D outline – we applied the following algorithms:
  - Support Vector Machine (SVM) [8]
  - Multi-Layer Perceptron (MLP) [9]
  - XGBoost (XGB) [10]
  - K-Nearest Neighbors (KNN) [11]
- Out of all the algorithms, SVM proved to be the most successful

Results

- Dataset of images consisting of 1,071 fish [7] across 7 genera, with each genus containing 153 specimens
- SVM classified fish with 96.3 ± 1.5% accuracy in 5-fold cross-validation
- Demonstrates that our EFD representation is distinguishable
- But it is not highly interpretable
    - High interpretability allows scientists to identify distinguishing characteristics of shapes with ease

Results

- Despite limitations, EFDs' ability to capture identifying data is evident in UMAP visualization [12]
    - Convert to HSV color space
- Each data point represents a fish specimen color-coded by genus
    - Clusters demonstrate the fish genera are well-separated in feature space



Conclusion

- We developed a computational method for automatically outlining specimens in museum image
- Outlines can be extracted from museum archives for use in scientific studies
- Research demonstrates promise of computational techniques to support researchers in leveraging and exploiting archival collections for scientific discoveries

Acknowledgements

This research was supported by Drexel University's STAR (Students Tackling Advanced Research) Scholars Program, and NSF grant OAC-1940233. We thank the Tulane University Biodiversity Research Institute (TUBRI) for their support, and Chris A. Taylor, Curator of Fishes and Crustaceans at the Illinois Natural History Survey (INHS). INHS is one of six fish collections participating in the Great Lakes Invasives Network (GLIN).

DREXEL UNIVERSITY College of Computing & Informatics

DREXEL UNIVERSITY Metadata Research Center College of Computing & Informatics

NSF

STAR SCHOLARS

References

1. G. Colavizza, T. Blanke, C. Jeurgens, and J. Noordegraaf, “Archives and AI: An overview of current debates and future perspectives,” Journal on Computing and Cultural Heritage, vol. 15, no. 1, pp. 4:1–15, 2021.
2. F. Kuhl and C. Giardina, “Elliptic Fourier features of a closed contour,” Computer Graphics and Image Processing, vol. 18, pp. 236–258, 1982
3. J. Pepper, A. Senin, D. Jebbia, D. Breen, and J. Greenberg, “Metadata verification: A workflow for computational archival science,” in Proc. IEEE International Conference on Big Data, 2022, pp. 2565– 2571.
4. K. Karnani, J. Pepper, Y. Bakis¸, X. Wang, H. Bart Jr., D. Breen, and J. Greenberg, “Computational metadata generation methods for biological specimen image collections,” International Journal on Digital Libraries, 2022. [Online]. Available: https://doi.org/10.1007/s00799- 022-00342-1
5. N. Otsu, “A threshold selection method from gray-level histograms,” IEEE Transactions on Systems, Man, and Cybernetics, vol. 9, no. 1, pp. 62–66, 1979.

References

6. Wu, Y.; Kirillov, A.; Massa, F.; Lo, W.Y.; Girshick, R.: Detectron2. https://github.com/facebookresearch/detectron2, 2019
7. Illinois Natural History Survey, “INHS Fish Collection,” https://fish.inhs.illinois.edu/, Accessed 11/2023.
8. V. Vapnik, “An overview of statistical learning theory,” IEEE Transactions on Neural Networks, vol. 10, no. 5, pp. 988–999, 1999.
9. D. Rumelhart, G. Hinton, and R. Williams, “Learning representations by back-propagating errors,” Nature, vol. 323, pp. 32–46, 1986.
10. J. H. Friedman, “Greedy function approximation: A gradient boosting machine,” The Annals of Statistics, vol. 29, no. 5, pp. 1189 – 1232, 2001.
11. T. Cover and P. Hart, “Nearest neighbor pattern classification,” IEEE Transactions on Information Theory, vol. 13, no. 1, pp. 21–27, 1967
12. L. McInnes, J. Healy, and J. Melville, “UMAP: Uniform manifold approximation and projection for dimension reduction,” 2020.

Questions