BISS Biodiversity Information Science and Standards

Biodiversity Information Science and Standards 7: e112742
doi: 10.3897/biss.7.112742

OPEN ACCESS

Conference Abstract

AI-Accelerated Digitisation of Insect Collections: The next generation of Angled Label Image Capture Equipment (ALICE)

Arianna Salili-James‡, Ben Scott‡, Laurence Livermore‡, Ben Price‡, Steen Dupont‡, Helen Hardy‡, Vincent Stuart Smith‡

‡ Natural History Museum, London, United Kingdom

Corresponding author: Vincent Stuart Smith (v.smith@nhm.ac.uk)

Received: 14 Sep 2023 | Published: 15 Sep 2023

Citation: Salili-James A, Scott B, Livermore L, Price B, Dupont S, Hardy H, Smith VS (2023) AI-Accelerated Digitisation of Insect Collections: The next generation of Angled Label Image Capture Equipment (ALICE). Biodiversity Information Science and Standards 7: e112742. https://doi.org/10.3897/biss.7.112742

Abstract

The digitisation of natural science specimens is a shared ambition of many of the largest collections, but the scale of these collections, estimated at at least 1.1 billion specimens (Johnson et al. 2023), continues to challenge even the most resource-rich organisations.

The Natural History Museum, London (NHM) has been pioneering work to accelerate the digitisation of its 80 million specimens. Since the inception of the NHM Digital Collection Programme in 2014, more than 5.5 million specimen records have been made digitally accessible. This has enabled the museum to deliver a tenfold increase in digitisation, compared to when rates were first measured by the NHM in 2008. Even with this investment, it will take circa 150 years to digitise its remaining collections, leading the museum to pursue technology-led solutions alongside increased funding to deliver the next increase in digitisation rate.

Insects comprise approximately half of all described species and, at the NHM, represent more than one-third (c. 30 million specimens) of the NHM's overall collection. Their most common preservation method, attached to a pin alongside a series of labels with metadata, makes insect specimens challenging to digitise. Early Artificial Intelligence (AI)-led innovations (Price et al. 2018) resulted in the development of ALICE, the museum's Angled

Label Image Capture Equipment, in which a pinned specimen is placed inside a multi-camera setup, which captures a series of partial views of a specimen and its labels. Centred around the pin, these images can be digitally combined and reconstructed, using the accompanying ALICE software, to provide a clean image of each label. To do this, a Convolutional Neural Network (CNN) model is incorporated, to locate all labels within the images. This is followed by various image processing tools to transform the labels into a two-dimensional viewpoint, align the associated label images together, and merge them into one label. This allows users to manually, or computationally (e.g., using Optical Character Recognition [OCR] tools) extract label data from the processed label images (Salili-James et al. 2022).

With the ALICE setup, a user might average imaging 800 digitised specimens per day, and exceptionally, up to 1,300. This compares with an average of 250 specimens or fewer daily, using more traditional methods involving separating the labels and photographing them off of the pin. Despite this, our original version of ALICE was only suited to a small subset of the collection. In situations when the specimen is very large, there are too many labels, or these labels are too close together, ALICE fails (Dupont and Price 2019).

Using a combination of updated AI processing tools, we hereby present ALICE version 2. This new version of ALICE provides faster rates, improved software accuracy, and a more streamlined pipeline. It includes the following updates:

- **Hardware**: after conducting various tests, we have optimised the camera setup. Further hardware updates include a Light-Emitting Diode (LED) ring light, as well as modifications to the camera mounting.
- **Software**: our latest software incorporates machine learning and other computer vision tools to segment labels from ALICE images and stitch them together more quickly and with a higher level of accuracy, significantly reducing the image processing failure rate. These processed label images can be combined with the latest OCR tools for automatic transcription and data segmentation.
- **Buildkit**: we aim to provide a toolkit that any individual or institution can incorporate into their digitisation pipeline. This includes hardware instructions, an extensive guide detailing the pipeline, and new software code accessible via Github.

We provide test data and workflows to demonstrate the potential of ALICE version 2 as an effective, accessible, and cost-saving solution to digitising pinned insect specimens. We also describe potential modifications, enabling it to work with other types of specimens.

Keywords

entomology, biodiversity, data, machine learning, computer vision

Presenting author

Vincent Stuart Smith

Presented at

TDWG 2023

Conflicts of interest

The authors have declared that no competing interests exist.

References

- Dupont S, Price B (2019) ALICE, MALICE and VILE: High throughput insect specimen digitisation using angled imaging techniques. Biodiversity Information Science and Standards 3 https://doi.org/10.3897/biss.3.37141
- Johnson K, Owens IP, Global Collections Group (2023) A global approach for natural history museum collections. Science 379 (6638): 1192-1194. https://doi.org/10.1126/science.adf6434
- Price BW, Dupont S, Allan EL, Blagoderov V, Butcher AJ, Durrant J, Holtzhausen P, Kokkini P, Livermore L, Hardy H, Smith V (2018) ALICE: Angled Label Image Capture and Extraction for high throughput insect specimen digitisation. OSF Preprints https://doi.org/10.31219/osf.io/s2p73
- Salili-James A, Scott B, Smith V (2022) ALICE Software: Machine learning & computer vision for automatic label extraction. Biodiversity Information Science and Standards 6 https://doi.org/10.3897/biss.6.91443