ALICE Software: Machine learning & computer vision for automatic label extraction

Arianna Salili-James‡, Ben Scott‡, Vincent S. Smith‡

‡ Natural History Museum, London, United Kingdom

Corresponding author: Arianna Salili-James (arianna.salili-james@nhm.ac.uk), Ben Scott (b.scott@nhm.ac.uk), Vincent S. Smith (v.smith@nhm.ac.uk)

Abstract

Insects make up over 70% of the world's known species (Resh and Carde 2009). This is well represented in collections across the world, with the Natural History Museum's pinned insect collection alone making up nearly 37% of the museum's remarkable 80 million specimen collection. Thus, this extraordinary dataset is a major focus of digitisation efforts here at the Museum. While hardware developments have seen digitisation processes significantly improve and speed up (Blagoderov et al. 2017), we now concentrate on the latest software and explore whether machine learning can lend a bigger hand in accelerating our digitisation of pinned insects.

Traditionally, the digitisation of pinned specimens involves the removal of labels (as well as any supplementary specimen miscellanies) prior to photographing the specimen. In order to document labels, this process is typically followed by additional photographs of labels as the label documentation is often obstructed by their stacking on a pin, the specimen and additional specimen material, or the pin itself. However, these steps not only slow down the process of digitisation but also increase the risk of specimen damage. This encouraged the team at the Natural History Museum to develop a novel setup that would bypass the need for removing labels during digitisation. This led to the development of ALICE (*Angled Label Image Capture and Extraction*) (Dupont and Price 2019).

ALICE is a multi-camera setup designed to capture images of angled specimens, which allows users to get a full picture of a specimen in a collection, including that of the label and the text within. Specifically, ALICE involves four cameras angled at different viewpoints in order to capture label information, as well as two additional cameras providing a lateral and dorsal view of the specimen. By viewing all the images taken from one specimen simultaneously, we can obtain a full account of the labels and of the specimen, despite any obstructions. This setup notably accelerates parts of the digitisation process, sometimes by up to 7 times (Price et al. 2019). Furthermore, ALICE presents the opportunity to incorporate machine learning and computer vision techniques to create a software that automates the process of transcribing the information contained on labels.

Automatically transcribing text (whether typed or handwritten) from label images, leads to the topic of Optical Character Recognition (OCR). Regardless of any obstructions to the labels, standard OCR methods will often fail at detecting text from these angled specimens if no preprocessing is done. This was emphasised in Bieniecki et al. (2007), which showed that some standard OCR algorithms were highly sensitive to geometric distortions such as bad angles. Therefore, in our latest ALICE software, we take on a 5-step approach that segments labels and merges them together using machine learning and computer vision, before turning to standard OCR tools to transcribe texts. Our 5-step framework is described as follows:

1. Label segmentation with Convolutional Neural Networks (CNNs),
2. Label corner approximation,
3. Perspective transformation of label, followed by image registration with a given template,
4. Label-image merging using an averaging technique,
5. OCR on merged label.

While ALICE aims to reveal specimen labels using a multi-camera setup, we ask ourselves whether an alternative approach can also be taken. This takes us to the next phase of our digitisation acceleration research on smarter cameras with cobots (collaborative robots) and the associated software. We explore the potential of a single camera setup that is capable of zooming into labels. Where intelligence was incorporated post-processing with ALICE, using cobots, we can incorporate machine learning and computer vision techniques in-situ, in order to extract label information. This all forms the focus of our presentation.

Keywords

digitisation, pinned specimens, label transcription, segmentation, cobots

Presenting author

Arianna Salili-James

Presented at

TDWG 2022

Conflicts of interest

References

- Bieniecki W, Grabowski S, Rozenberg W (2007) Image Preprocessing for Improving OCR Accuracy. In: Bieniecki W (Ed.) 2007 international conference on perspective technologies and methods in MEMS design. https://doi.org/10.1109/MEMSTECH.2007.4283429
- Blagoderov V, Penn M, Sadka M, Hine A, Brooks S, Siebert D, Sleep C, Cafferty S, Cane E, Martin G, Toloni F, Wing P, Chainey J, Duffell L, Huxley R, Ledger S, McLaughlin C, Mazzetta G, Perera J, Crowther R, Douglas L, Durant J, Honey M, Huertas B, Howard T, Carter V, Albuquerque S, Paterson G, Kitching I (2017) *iCollections* methodology: workflow, results and lessons learned. Biodiversity data journal https://doi.org/10.3897/bdj.5.e19893
- 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
- Price B, Dupont S, Allan E, Kokkini P, Blagoderov V, Butcher A, Durrant J, Holtzhausen P, Livermore L, Hardy H, Smith V (2019) ALICE: Angled Label Image Capture and Extraction for high throughput insect specimen digitisation. OSF Preprints https://doi.org/10.31219/osf.io/s2p73
- Resh VH, Carde RT (2009) Encyclopedia of Insects. Academic press https://doi.org/10.1016/B978-0-12-374144-8.X0001-X