Quantifying Bird Skeletons

Zhizhuo Zhou\textsuperscript{1}, Gemmechu Hassena\textsuperscript{2}, Brian C. Weeks\textsuperscript{3}, David F. Fouhey\textsuperscript{1}
\textsuperscript{1} University of Michigan EECS, \textsuperscript{2} Addis Ababa University, \textsuperscript{3} University of Michigan SEAS

Abstract

Birds have well-known distributions, phylogenetic relationships, and life histories, making them a powerful model system for understanding biotic responses to global environmental change. However, there are hundreds of thousands of museum skeletal specimens that could be analyzed to further our understanding of avian responses to climate change, but remain under-utilized due to the practical constraints of measuring elements of skeletons by hand. We introduce a dataset and system for measuring skeletal traits from museum specimens that reduces capture time by 15x for an initial effort to measure 10 traits per skeleton, allows for post hoc addition of trait data orders of magnitude faster than traditional methods, and shows high accuracy even when trained with limited data.

1. Introduction and Related Work

Climate change is predicted to result in changes in animal body size, with warming temperatures expected to result in size reductions broadly across the tree of life [8]. Size impacts nearly every aspect of an organism’s ecology, thus understanding how species compensate for warming-driven size reductions is key to understanding and predicting the impacts of climate change on the world’s ecosystems.

In birds, recent warming has driven consistent reductions in body size, with contemporaneous changes in body shape [19]. While it is generally known that morphology is predictive of ecology [16], key to understanding the impacts of climate change on animal ecology and fitness is the collection of extensive and detailed functional trait data from a large number of individuals within species through time. Museum skeletal specimens could fill this data gap: over 500,000 bird skeletal specimens are held in natural history museums. However, measuring only a limited number of traits from a specimen can take a well-trained technician 15-30 minutes; as a result, collecting only a small number of species by hand at the timescales necessary to understand biotic responses to climate change can take decades [19].

We aim to automate the measurement of bone lengths from a single, casually arranged, top-down image of a bird skeleton specimen. This casual capture (Figure 1) enables rapid digitization (taking 1 minute), but introduces challenges: bones intersect and appear in arbitrary rotations, and bone categories must be recognized across varying species. As a step towards digitizing a larger ≈25K skeleton collection, we introduce an annotated dataset in Section 2 that includes both pixel annotations and hand-measurements. We describe a measurement system based on Mask-RCNN [9] and synthetic augmentation [7] in Section 3, which we train on our preliminary data.

Our experiments (Section 4) show that our system, when trained on 150 preliminary specimens, automatically measures 60% of bones within 5% relative error. Some types (e.g., Humerus) are measured with 90% within 5% relative error and many bones are measured to sub-mm precision. We hope our method will enable large-scale ecological studies and contribute to our understanding of the links between climate change and ecology.

Related Work: Our work is part of a broader trend of using deep-learning based tools to solve research questions in the sciences. This has been applied to topics ranging from botany [15, 18] to keypoint detection [14, 12] and or 3D reconstruction [4, 2, 3, 21]. While our animal specimens
do not move, making measurement easier, our measurements need high precision, which leads to a special capture setup. Our annotations require extensive expert time, which shapes data collection and leads to heterogeneously-annotated data, which we handle by self-training.

We frame our problem via instance segmentation to take advantage of the well-developed segmentation tools in the community [9, 20, 11]. In our work, we use Mask R-CNN [9] for our method, and the Toronto Annotation Suite [1] to accelerate our annotation. We similarly take inspiration from improvements to instance segmentation using synthetic data [7]. Building on this well-developed toolchain enables our approach to obtain promising performance with a fairly small amount of annotation.

2. Dataset

We evaluate methods on a 376 annotated image subset out of a larger in-progress effort that has captured over 10,000 bird skeleton images. Of the 376 images, 250 are annotated with pixel annotations. We hold out another 126 that are separately hand-measured with bone lengths; the non-unified set stems from Covid-19 restrictions. For now, we focus on 10 Bones of Interest, labeling other bones as other. We collect our dataset in four stages. All images are Captured in a process that is rapid and does not require specialized placement. Our annotated data is then processed with multiple stages to maximize use of experts’ time, consisting of a Bone Spotting done by a trained technician, and Pixel Annotation that can often done by a non-trained expert. Finally, for empirical validation, we also do Physical measurement annotation of a small number of specimens.

**Image Capture:** Capturing a top-down image of a specimen entails retrieving the specimen and spreading the bones onto the capture surface. This is done casually to ensure the process takes only around a minute. These bones are imaged with a FLIR Blackfly S camera with a SONY IMX183 sensor ≈400mm above the center of the capture surface. In setup, 1px error corresponds to ≈0.07mm error.

**Bone Spotting:** Pixel-annotation of all bones of interest takes a long time, so to enable parallelizability and efficiently use expert time, we first complete a bone spotting step. In this step, the expert provides a rough bounding polygon and name of all bones. This enables parallelizing the slower segmentation task. We use the VGG Image Annotator [6] to annotate bounding polygons.

**Pixel Annotations:** Once bones of interest have been separated, we annotate cropped images around the bounding polygons. Easy bones (e.g., humerus) can be annotated by non-experts. We use the Toronto Annotation Suite [1], powered by human-in-the-loop AI, which substantially accelerates annotation. While all bones have a bounding polygon from the expert annotation stage, “other” and broken ones do not have pixel segmentations. This creates a mixed dataset of bounding polygons and pixel-wise segmentations, shown in Figure 2. We describe our approach to handle the mixed annotations in Section 3.

**Physical Measurement:** To provide end-to-end validation of the full system in terms of measurement in physical quantities, we hand-measure a held-out test set of 126 specimens. We hand measure the length of each bone of the specimen and record it to the nearest hundredth of a mm.

3. Method

We build a skeleton measurement system to compute physical measurements of the bones of interest from an image. Our core system is built on MaskRCNN [9]. We perform additional steps to enable best use of the data that is available: there are many “other” bones without pixel annotation, and the number of bones of interest with pixel segmentation is small. We produce a homogeneous dataset by using noisy self-training to annotate the other bones and synthetic data to expand our dataset.
Table 1. AP on a test set of 50 images with Mask R-CNN trained on different datasets described in Section 3.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Carpo-metacarpus</th>
<th>Femur</th>
<th>Humerus</th>
<th>Radius</th>
<th>Sclerotic Ring</th>
<th>Second Digit</th>
<th>Skull</th>
<th>Tarso-metatarsus</th>
<th>Tibiotarsus</th>
<th>Ulna</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original</td>
<td>46.0</td>
<td>44.7</td>
<td>51.3</td>
<td>71.5</td>
<td>15.2</td>
<td>75.2</td>
<td>17.9</td>
<td>91.3</td>
<td>42.2</td>
<td>32.0</td>
</tr>
<tr>
<td>Noisy</td>
<td>52.2</td>
<td>47.7</td>
<td>55.4</td>
<td>72.5</td>
<td>18.4</td>
<td>76.2</td>
<td>23.1</td>
<td>92.1</td>
<td>41.1</td>
<td>57.5</td>
</tr>
<tr>
<td>Synthetic</td>
<td>38.1</td>
<td>25.9</td>
<td>41.7</td>
<td>61.9</td>
<td>7.6</td>
<td>64.0</td>
<td>8.0</td>
<td>81.6</td>
<td>18.9</td>
<td>38.8</td>
</tr>
<tr>
<td>Noisy + Synthetic</td>
<td><strong>57.1</strong></td>
<td><strong>51.1</strong></td>
<td><strong>62.3</strong></td>
<td><strong>72.4</strong></td>
<td><strong>25.3</strong></td>
<td><strong>78.3</strong></td>
<td><strong>34.9</strong></td>
<td><strong>92.4</strong></td>
<td><strong>46.8</strong></td>
<td><strong>56.9</strong></td>
</tr>
</tbody>
</table>

Figure 3. Left: a sample instance segmentation output with extracted and reoriented bones on the bottom. Right: a showcase for randomly selected carpometacarpus, humerus, and skull segmentations, demonstrating the potential of our system for large-scale skeleton analysis.

**Skeleton Measurement System:** The skeleton measurement system comprises an instance segmentation system which identifies the type of bone and its segmentation, followed by simple image processing to produce annotations. Specifically, we train Mask R-CNN [9] to predict bounding boxes plus class labels (the 10 bones of interest, plus other) as well as segmentation masks. Once these bones have been segmented, we extract the longest diagonal of the segmentation mask. This pixel measurement is converted to millimeters by assuming the bone is on the surface and that the surface and the image plane are parallel.

**Noisy Annotation Generation:** Since the “other bones” bones are spotted but not segmented, we create a homogeneous dataset by self-training. Given the initial annotations for the bones of interest, we train a binary U-Net [17] that predicts bone-vs-background; we minimize loss only on regions that have been pixel-annotated (i.e., ignoring other bones). The resulting annotation is then merged with the original dataset to produce what we call a Noisy Dataset.

**Synthetic Data Generation:** Finally, we expand our training set by building a synthetic dataset following [7]. This helps get additional performance out of the limited annotations that can be generated. We generate images that mimic the data capture process by: (1) picking a background; (2) picking a random sample of segmented bones from the training set that models the skeletal composition of a bird; and (3) randomly selecting a location for each bone from a list of anchor locations. Finally, we composite the image with randomized blending parameters (Gaussian blur kernel size, and erosion kernel), similar to [7]. Parameter ranges were tuned manually to make the resulting images realistic. We refer to the results as the Synthetic Dataset.

**Implementation Details:** Our system uses Mask R-CNN [9] model with ResNet50 [10] and FPN [13] backbone using pre-trained weights from [20]. We train all models for 12,000 steps using Adam with base learning rate of .001 and the default Detectron2 [20] hyperparameters. We train with a resolution of 1368×912 pixels (4× smaller) for memory usage, making a pixel ≈0.28mm. Our self-training U-Net [17] has 4 max pooling and 4 up-sampling blocks, which each contain 2 convolutional layers with kernel size 3.

4. Experiments

We now evaluate the performance of our bone measurement system. We evaluate the system in two modes – pixel-based metrics that test the system’s ability to segment bones of interest, and end-to-end metrics that test its ability to measure bones in millimeters. We show some examples of our system segmenting objects in Figure 3, including on a single image as well as showing a collection of segmented
We compare Mask R-CNN-based Evaluation:
detected and measured within 5% relative error.

Considering 1k and 2k synthetic images. If there is too much
duplicate real data during training, peak performance is be-
due to overfitting to artifacts. Table 1 shows an extreme case where only
synthetic data has poor performance.

End-to-End Evaluation: We finally evaluate the end-to-
end performance by analyzing the percent of bones that are
detected and measured within 5% relative error across dif-
ferent classes. Table 2 shows a similar trend: noisy and
synthetic data aids performance. Gains are not as strong as in
Table 1, suggesting that segmentation accuracy is corre-
lated, but only partially so, with measurement performance.

To give a sense of the scale of bones, we also report the
mean error in Table 3, computed only on bones detected
by all three models in comparison. This comparison is
needed to ensure apples-to-apples comparisons across mod-
els. Many average errors for bones are sub-millimeter (e.g.,
sclerotic ring), enabled by both highly accurate segmenta-
tion models and high-quality cameras. Even after scaling
by the relatively small bones, these errors are small. For in-
stance, the average humerus error (0.52mm) is 2.3% of the
length of the average humerus (22.7mm). Again, increasing
synthetic data improves results, but not as dramatically as in
segmentation.

5. Conclusion

We presented a system for skeletal trait measurement.
Our preliminary results suggest that the current pipeline can
can obtain highly accurate results even with limited data.
We expect that increasing the number of real images in use
will substantially improve results. We believe our work will
enable us to tap data on the relationship between climate
and animal ecology on an unprecedented scale. Future work
includes expanding bones of interest and increasing seg-
mentation resolution.

Acknowledgments: This work is based on specimens from
the University of Michigan Museum of Zoology, Depart-
ment of Ornithology, with the support of Benjamin Winger,
Curator of Birds, and Brett Benz, Collection Manager, Di-
vision of Ornithology.
References

[21] Silvia Zuffi, Angjoo Kanazawa, David Jacobs, and Michael J. Black. 3D menagerie: Modeling the 3D shape and pose of animals. In CVPR, July 2017. 1