Zero-Shot Learning and Detection of Teeth in Images of Bat Skulls

Xu Hu, Michael Lam, Sinisa Todorovic, Thomas G. Dietterich
EECS Department, Oregon State University
Corvallis, Oregon, 97331-5501
{huxu,lamm,sinisa,tgd}@eecs.oregonstate.edu

Maureen A. O'Leary
Department of Anatomical Sciences, HSC T-8 (040), Stony Brook University
Stony Brook, New York, 11794-8081
maureen.oleary@stonybrook.edu

Andrea L. Cirranello, Nancy B. Simmons, Paúl M. Velazco
Department of Mammalogy, American Museum of Natural History
79th Street and Central Park West, New York, New York, 10024-5192
{awetterer,simmons,pvelazco}@amnh.org

Abstract

*Biologists collect and analyze phenomic (*e.g.*, anatomical or non-genomic) data to discover relationships among species in the Tree of Life. The domain is seeking to modernize this very time-consuming and largely manual process. We have developed an approach to detect and localize object parts in standardized images of bat skulls. This approach has been further developed for unannotated images by leveraging knowledge learned from a few annotated images. The key challenge is that the unlabeled images show bat skulls of "unknown" species that may have types, total numbers, and layouts of the teeth that differ from the "known" species appearing in the labeled images. Our method begins by matching the unlabeled images to the labeled ones. This allows a transfer of tooth annotations to the unlabeled images. We then learn a tree parts model on the transferred annotations, and apply this model to detect and label teeth in the unlabeled images. Our evaluation demonstrates good performance, which is close to our upper bound performance by the fully supervised model.*

1. Introduction

"Phenomic characters" represent a rich source of information for understanding biodiversity and evolution [15, 5], especially for reconstructing the Tree of Life. For fossil species, these data are the only way to discover the evolutionary relationships among species. For living species, phenomic data contribute to our understanding of evolutionary relationships and provide a window into the complex interrelationships of form, environment, and genes. Phenomic characters include anatomical characteristics of organisms, such as presence or absence of shared or unique parts (*e.g.*, horns, wings), shapes of parts (coiled versus straight horns), relationships between parts (*e.g.*, that the eye is superior to the nose), and other features such as biochemistry and behavior. MorphoBank, a new web application and database allows researchers to collect and archive images collaboratively in online matrices [8, 9]. MorphoBank now includes thousands of scores and annotated images used in evolutionary research. Columns in MorphoBank matrices represent characters, such as the presence/absence of a part (*e.g.*, horns) or more complex relationships (*e.g.*, distance between teeth). Rows in matrices are species. Scoring each cell in a matrix, however, currently requires individual visual inspection by an expert, limiting the speed at which these data can be analyzed. A goal of our research, the collaborative AVAToL [1] project, is to apply computer vision to accelerate this process, which will significantly advance the reconstruction of the whole Tree of Life containing tens of millions of species [16].

Here, we use an image collection of three standardized views of skulls (dorsal, ventral, and lateral anatomical orientations) of eight species of bats (Chiroptera, Mammalia) provided by researchers in the Department of Mammalogy at the American Museum of Natural History. We attempt

[1] NSF - Assembling, Visualizing, and Analyzing the Tree of Life; http://avatol.org/ngp/

to score the types and layout of teeth. Teeth vary widely in mammalian evolution and their differences are important distinguishing characteristics of species and larger groups. Mammals, including bats, have four tooth types (incisors, canines, premolars, and molars). Differentiating tooth type and number can be especially problematic when teeth appear to be similar in texture and shape.

Few skull images show the ventral (bottom) or lateral (side) view of a bat skull against an uniform background, as illustrated in Figures 1 and 2. Note that teeth only show up in these two views, thus only images of ventral and lateral views are used in our experiments. Since manual annotations are expensive, only a subset of images of a few bat species were annotated with the locations and types of teeth present. We refer to these species as "known bat species". Note that in our experiments, the image acquisition process readily provides metadata about the species and view (ventral or lateral) of each specimen/image. Thus, in this work, image annotation does not pertain to the class label and view of an object occurring in the image, as is typically the case in the object recognition literature, but to the bounding boxes (and associated part name) drawn around each object part of interest in the image. Our goal is to detect and localize the teeth present in the much more numerous unannotated images of other "unknown" bat species. They are "unknown" species, because we do not know a priori the locations and types of teeth present in the unannotated images.

This is a very challenging problem for state-of-the-art computer vision. First, fine-grained recognition is required, because certain types of adjacent teeth on a bat skull appear very similar. Second, tooth localization must be highly accurate, since measuring relative displacement is important for biological studies. These challenges could potentially be addressed by leveraging contextual and spatial constraints (*e.g.*, premolars are located between canine and molar teeth). However, in our setting, contextual reasoning is very difficult, because the eight collected bat species have totally different tooth numbers, spatial layouts, and types of teeth (*e.g.*, some species lack all but a single premolar). One could learn these phenomic characters for the known bat species in the annotated images. But, for the remaining unknown bat species, learning the numbers, configurations, and types of teeth must be accomplished without any training examples or supervision, *i.e.*, zero-shot learning.

Because the known and unknown bat species share many phenomic characters, it is reasonable to expect that transferring knowledge about the known species would enable robust zero-shot learning of the unknown species. This, in turn, would allow successful tooth detection and localization in the unannotated images. Transfer learning provides a framework to utilize prior knowledge so that a model of a new class can be trained on only a few training examples and yet match the performance attained by standard training on many training examples. Related work typically focuses on transfer learning for image classification. For example, transfer learning can be used to find a feature representation that is shared by all image classes to address the lack of training data for some image classes [12, 11]. Similarly, learning shared scene parts [18], and shared training examples [6] has been shown to produce successful transfer learning in image classification. Transfer learning has also been used for attribute recognition by finding a shared classifier of object attributes [1, 14]. Zero-shot learning has been achieved for image classification by mapping the raw features of the unknown class into a common feature space of the known classes [4]. However, most of these approaches cannot be easily extended to address object detection and localization. We are not aware of any approach to zero-shot learning that is able to estimate unknown numbers, layouts, and types of parts of new objects.

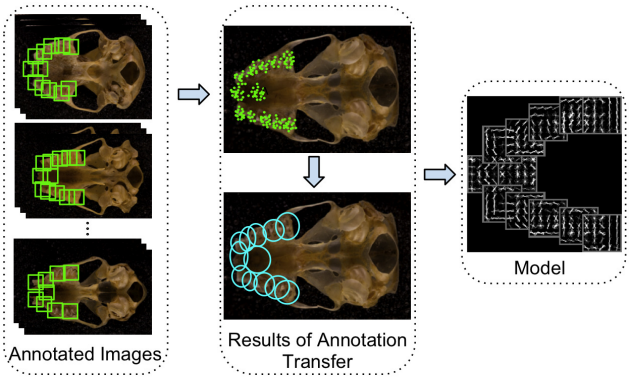


Figure 1: **Left**: Annotated images of known bat species. **Middle**: Annotated bounding boxes of known bat species are transferred to the unannotated images by image matching. Distributions of transferred annotations are estimated (shown as circles). **Right**: A tree-structured part model for the unknown species is learned from the distributions of transferred part annotations.

In this work, we make the assumption that the set of annotated images is sufficiently large to provide information about all the teeth types, so that no new types are expected in the unannotated images. The teeth types include: incisors (labeled I1), canines (C), premolars (P4, P5), and molars (M1, M2). Since bat skulls exhibit approximate axial symmetry, we expect that if a specimen has a particular tooth type, then a pair of teeth of that type occur symmetrically on each side of the jaw. Since the labels of species and view (ventral or lateral) are given for every unannotated specimen/image, we formulate zero-shot learning for the case when the considered set of unannotated images show the same view of skulls of only one species.

Our approach is illustrated in Fig. 1. First, we match the unannotated images to the annotated ones. This allows a transfer of tooth annotations to the unannotated images. Since image matching is subject to noise, for every tooth type, we estimate a mixture of Gaussian distributions of locations of matches in the unannotated images. Second, the Gaussian mixtures are used to robustly learn a part-based model of the new bat species. The layout of parts is modeled as a tree representing the symmetric arrangement of the teeth along the jaw. The tree root represents the upper mesial incisors. Third, to identify the total number of teeth in the model, *i.e.*, to perform model identification, we start with the tree model which has all the tooth types. Then, we conduct hypothesis testing to sequentially remove parts from the tree model (and their symmetric pairs) whose displacements relative to neighboring parts are significantly different from those observed in the annotated images. The resulting tree model is then employed for ultimate tooth detection and localization. In this work, we do not integrate the two tree models learned on ventral and lateral views of a given "unknown" bat species. Instead, we treat the models as two independent representations of the object class.

Contributions. To the best of our knowledge, this paper presents the first approach to zero-shot learning for part detection and localization. The related work of [7] uses dense image matching to transfer labels of segments from annotated images, but for the purpose of improving segmentation of unannotated images — not for part detection and localization. We formulate a novel soft loss for the structured output learning of our tree model.

In the following, Sec. 2–4 specifies the tree model and its inference and learning; Sec. 5 explains our annotation transfer; Sec. 6 describes model identification; and Sec. 7 presents the dataset of bat skull images and our results.

2. The Tree Model

This section specifies our tree model for detection and localization of object parts. Let $\mathbb{D} = \{I_m : m = 1, 2, \dots\}$ denote the set of our bat skull images (all assumed to have the same size of $H \times W$ pixels). Each image shows either the ventral view or lateral view of a bat skull with the teeth $\mathcal{V} = \{\mathcal{V}_i : i = 1, \dots, n\}$, where $\mathcal{V}_i$ is an instance of a particular tooth type. The teeth occur in an image in a particular symmetric configuration, as illustrated in Fig. 1, which can be modeled as a directed tree, $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{E}$ is the set of edges connecting the neighboring teeth (*i.e.*, nodes). As the tree root, we specify a part of the bat skull that is assumed always present in all the images. In particular, the root represents the upper mesial incisors in the skull. Each tooth $\mathcal{V}_i \in \mathcal{V}$ is characterized by location $\mathbf{s}_i \in \mathcal{S}_i = \{1, \cdots, H\} \times \{1, \cdots, W\}$ in the image.

We define structured output space $\mathcal{S} = \mathcal{S}_1 \times \cdots \times \mathcal{S}_{|\mathcal{V}|}$, and cast part detection and localization as a structured prediction problem. Specifically, the true teeth locations $S = \{\mathbf{s}_i : i = 1, \dots, n\}$ are assumed to maximize the score, $f(I, S)$, defined as

$$f(I, S) = \sum_{\mathcal{V}_i \in \mathcal{V}} \theta_i(I, \mathbf{s}_i) + \sum_{(\mathcal{V}_i, \mathcal{V}_j) \in \mathcal{E}} \theta_{ij}(\mathbf{s}_i, \mathbf{s}_j), \quad (1)$$

where $\theta_i(I, \mathbf{s}_i)$ is the appearance score, and $\theta_{ij}(\mathbf{s}_i, \mathbf{s}_j)$ is the deformation score, as in the pictorial structure and deformable parts models [2, 3]. Thus, part detection and localization amounts to estimating $\hat{S} = \arg\max_{S \in \mathcal{S}} f(I, S)$.

We specify $\theta_i(I, \mathbf{s}_i)$ and $\theta_{ij}(\mathbf{s}_i, \mathbf{s}_j)$ as linear functions:

$$\theta_i(I, \mathbf{s}_i) = \langle \mathbf{w}_i, \boldsymbol{\psi}_i(I, \mathbf{s}_i) \rangle, \quad (2)$$

$$\theta_{ij}(\mathbf{s}_i, \mathbf{s}_j) = \langle \mathbf{w}_{ij}, \boldsymbol{\psi}_{ij}(I, \mathbf{s}_i, \mathbf{s}_j) \rangle, \quad (3)$$

where $\boldsymbol{\psi}_i$ is a feature descriptor vector associated with $\mathcal{V}_i$, and $\boldsymbol{\psi}_{ij}$ is a pairwise feature vector.

In this paper, $\boldsymbol{\psi}_i(I, \mathbf{s}_i)$ represents the standard HOG (histogram of oriented gradients) descriptor extracted from a window centered at $\mathbf{s}_i$. $\boldsymbol{\psi}_{ij}(I, \mathbf{s}_i, \mathbf{s}_j)$ is defined as the standard vector of displacements between windows centered at $\mathbf{s}_i$ and $\mathbf{s}_j$, $\boldsymbol{\psi}_{ij}(I, \mathbf{s}_i, \mathbf{s}_j) = [dx, dy, dx^2, dy^2]$.

All $\boldsymbol{\psi}_i(I, \mathbf{s}_i)$ and $\boldsymbol{\psi}_{ij}(I, \mathbf{s}_i, \mathbf{s}_j)$ are concatenated to form a joint feature map, $\boldsymbol{\Psi} = [\cdots, \boldsymbol{\psi}_i, \cdots, \boldsymbol{\psi}_{ij}, \cdots]$. The corresponding parameter vector is $\mathbf{w} = [\cdots, \mathbf{w}_i, \cdots, \mathbf{w}_{ij}, \cdots]$. From (1), the prediction score is $f(I, S) = \langle \mathbf{w}, \boldsymbol{\Psi}(I, S) \rangle$.

3. Inference

For the tree model defined in Sec. 2, we perform inference, $\hat{S} = \arg\max_{S \in \mathcal{S}} f(I, S)$, over $H \times W$ possible coordinates for each part $\mathcal{V}_i$, using the well-known generalized distance transform algorithm [2]. The complexity of our inference is $\mathbb{O}(H + W)$.

4. Learning

This section explains how to learn the parameter vector, $\mathbf{w}$, of our tree model from the unannotated images in the dataset. We formulate zero-shot learning as regularized risk minimization. Specifically, $\mathbf{w}$ is learned by the structured output SVM (SOSVM) [13]. SOSVM requires: 1) User-defined loss function $\Delta : \mathcal{S} \times \mathcal{S} \to \mathcal{R}$; and 2) Loss augmented inference, as specified below.

First, recall that we do not have access to ground-truth locations of the teeth $S = \{\mathbf{s}_i\}$. As mentioned in Sec. 1, we could only transfer annotations of tooth locations from the annotated images to the unannotated ones. Therefore, we define $\Delta(S, \hat{S})$ as an average of distances $K(\mathbf{s}_i, \hat{\mathbf{s}}_i)$ between the transferred annotations and predictions for each tooth as

$$\Delta(S, \hat{S}) = \frac{\gamma}{|\mathcal{V}|} \sum_{i=1}^{|\mathcal{V}|} K(\mathbf{s}_i, \hat{\mathbf{s}}_i), \quad (4)$$

where γ is a constant. For addressing noise in the annotation transfer along with meeting the domain requirements for highly accurate tooth location estimation, K is specified as a mixture of truncated Euclidean distances:

$$K(\mathbf{s}_i, \hat{\mathbf{s}}_i) = \sum_k \pi_k \left[(\boldsymbol{\mu}_{i;k} - \hat{\mathbf{s}}_i)^{\mathrm{T}} \Sigma_{i;k}^{-1} (\boldsymbol{\mu}_{i;k} - \hat{\mathbf{s}}_i)^{\mathrm{T}} \right]_\eta , \quad (5)$$

where $\pi_{i;k}$ are mixing parameters, and $\boldsymbol{\mu}_{i;k}$ and $\Sigma_{i;k}$ denote the kth mean and covariance of locations of the ith tooth transferred from the annotated images (see Fig. 1). Also, in (5), η is the minimum tolerable error in estimating part locations, such that $[c]_\eta = c$, if $c > \eta$, and $[c]_\eta = 0$, otherwise.

Note that in the special case, when $k = 1$, $\eta = 0$, and Σ is the identity matrix, K becomes the Euclidean distance. A similar truncated Euclidean loss was used for training part locations in [3]. It is also worth noting that our loss $\Delta(S, \hat{S})$ is different from that used for learning a Deformable Parts Model (DPM) of human poses, presented in [17]. Their loss accounts only for root predictions. In contrast, our learning controls performance of part localization by directly penalizing all individual incorrect part predictions.

Second, following the formulation of SOSVM [13], our learning requires loss-augmented inference of hidden teeth locations in the unannotated images, $h(I_m, S_m)$, which is defined as

$$S^* = \arg\max_{\hat{S}} \left[\Delta(S, \hat{S}) + f(I, \hat{S}) \right], \quad (6)$$

where S denotes the noisy teeth annotations transferred to image I, and $\hat{S}$ denotes the estimated teeth locations. Thus, the learning objective $F(\mathbf{w})$ is given by

$$\min_{\mathbf{w}} \frac{\lambda}{2} \|\mathbf{w}\|_2^2 + \sum_m \max(0, h(I_m, S_m) - f(I_m, S_m)). \quad (7)$$

where λ is a constant, and the second term represents the surrogate hinge loss, *i.e.*, a convex upper bound of $\Delta(S, \hat{S})$.

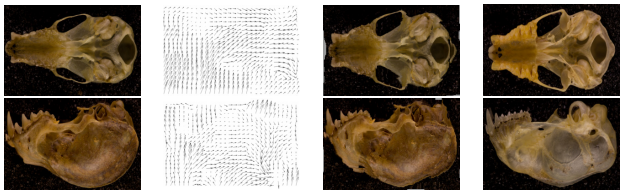
As in [10], we solve (7) using subgradient descent. By Danskin's theorem, a subgradient of the loss term in (7) is equal to $(\boldsymbol{\Psi}(I_m, S_m^*) - \boldsymbol{\Psi}(I_m, S_m))$. Thus, the subgradient of the objective function in (7) is

$$\frac{\partial F}{\partial \mathbf{w}} = \lambda \mathbf{w} + \sum_m (\boldsymbol{\Psi}(I_m, S_m^*) - \boldsymbol{\Psi}(I_m, S_m)). \quad (8)$$

For computing $\boldsymbol{\Psi}(I_m, S_m)$, we estimate the mean locations of the transferred annotations S_m and extract their HOG features. We obtain the model parameters $\mathbf{w}$ by performing gradient descent.

5. Annotation Transfer

This section explains how to transfer available part annotations of known objects to the unlabeled images. To this end, we use dense image matching. Matching pixels of the annotated images with pixels of the unannotated ones allows a transfer of tooth annotations from the set of annotated images. Note that, in our case, differences in the skulls of distinct bat species occurring in the images give rise to matching errors, even in the case of ideal matching. These errors in transferring tooth annotations could negatively affect our zero-shot learning. Consequently, we transfer tooth annotations only from a subset of known bat species whose image matching is estimated to be successful. Thus, our annotation transfer consists of two steps. We first match all annotated images of a known bat species to the set of unannotated images. Then we make a decision whether to use the matching results based on the estimated quality of the matching. Below, we formulate dense image matching.



(a) (b) (c) (d)

Figure 2: Matching: (a) Source images; (b) Resulting displacement fields; (c) Source images warped by the displacement fields to look like target images; (d) Target images.

Let $\mathbf{x} = [x, y]$ denote a pixel's coordinate in the source image I. For every pixel in I, we seek displacement $\mathbf{d}_\mathbf{x}$, such that $\mathbf{x} + \mathbf{d}_\mathbf{x}$ is that pixel's corresponding location in the target image I'. Thus, the matching between I and I' can be defined as an energy minimization problem:

$$\min_{\mathbf{d}} \sum_{\mathbf{x} \in I} \Big[\|\mathbf{g}(I(\mathbf{x})) - \mathbf{g}(I'(\mathbf{x} + \mathbf{d}_\mathbf{x}))\| + \beta \|\mathbf{d}_\mathbf{x}\| + \nu \sum_{\mathbf{y} \in N(\mathbf{x})} \min(\|\mathbf{d}_\mathbf{x}(x) - \mathbf{d}_\mathbf{y}(x)\|, \xi) \Big] \quad (9)$$

where $\mathbf{d} = \{\mathbf{d}_\mathbf{x} : \mathbf{x} \in I\}$ is the displacement field, $\mathbf{g}(I(\mathbf{x}))$ is the HOG descriptor extracted from image I at pixel location $\mathbf{x}$, $N(\mathbf{x})$ is the set of neighoring pixels of $\mathbf{x}$, and β, ν, ξ are constants. The first term in (9) penalizes large appearance differences between the candidate locations in I and I' for matching. The second regularization term in (9) models our domain knowledge that I and I' are generally similar, and thus the matching displacements should not be large. Finally, the last smoothness term in (9) penalizes matches that are inconsistent with the matches of neighbors. A similar formulation of dense image matching is presented in [7]. We solve (9) using tree-weighted max product message passing [7]. Examples of the resulting displacement fields are shown in Figure 2.

In the second step of our annotation transfer, we esti-

mate the quality of image matching. We expect that large errors in matching can be reliably detected by analyzing the resulting energy in (9). We sum the energy values of all image pairs of a given known species and the unknown species and then exclude the known species from annotation transfer when the energy sum is above a threshold. In our experiments, we adaptively set this threshold to remove from annotation transfer the one known species with the lowest matching energy.

6. Model Identification

As described in Sec. 4, we initially learn an all-teeth model of the unknown bat species. This section describes how to estimate the correct number of teeth in the model. We first detect tooth instances in the unannotated images using the all-teeth model, as described in Sec. 3. Then, we use these detections to identify particular tooth types to be removed from or kept in the model.

We expect that the tooth detections are placed in the images at random distances from one another, where the relative tooth-tooth distances in all the images are governed by an unknown distribution. Consequently, identifying whether a particular tooth type is present in the unknown bat species can be performed as a statistical hypothesis test of distances between the corresponding tooth detections in the unannotated images. We use the following two-sample t-test.

Let $P = \{p_l : l = 1, \ldots, L\}$ denote a set of all tooth types that appear in at least one bat species. For a particular tooth type, p_l, the null hypothesis H_0 states that the estimated distances in the unannotated images between p_l and its two nearest neighbors $N(p_l) \subset P$ come from the same underlying distribution as the corresponding distances in the annotated images. The alternative hypothesis H_1 states otherwise. We set the significance level to 5%, which means that we are confident that H_0 is true if p-value is greater than 0.05. If the t-test fails for a particular p_l, we remove p_l from the model and directly connect its two nearest neighbors $N(p_l) \subset P$ in the tree model. We keep performing the t-test until all the teeth in P are examined.

7. Experiments

The image dataset used for evaluation in this work has been collected by researchers in the Department of Mammalogy at American Museum of Natural History. The dataset consists of 160×2 images of 160 different specimens of bat skulls, each imaged from the ventral and lateral views. While the specimens are placed on a relatively uniform background, detecting certain skull parts of interest (in our case the teeth) is challenging due to their low contrast and very subtle differences in their shapes and textures. The specimens belong to 8 bat species, where each species is represented by 20×2 images. The eight species include: *Artibeus jamaicensis*, *Desmodus rotundus*, *Glossophaga soricina*, *Noctilio albiventris*, *Molossus molossus*, *Mormoops megalophylla*, *Saccopteryx bilineata* and *Trachops cirrhosus*. Each of these species has a subset of the following teeth: incisors (labeled I1), canines (C), premolars (P4, P5), and molars (M1, M2). Table 1 summarizes the tooth presence/absence characters for each species. Teeth of the same type differ in shape, texture, and layout across the eight bat species.

In our experiments, we use the "leave-one-out" (LOO) setting, where one bat species is treated as "unknown", and the remaining seven bat species are considered "known". Note that ground-truth tooth annotations of the selected "unknown" species are used only for evaluation, not for our zero-shot learning. For the "unknown" bat species, we zero-shot-learn two independent models on the ventral and lateral views. Then, the two models are used to detect and localize the tooth types in the corresponding views. The zero-shot learning is conducted on a subset of 12 randomly selected images per species. Our evaluation of detection and localization is performed on the remaining 8 images. Below, we first present our detection results, then, dense matching performance, and, finally, localization results. All results are averaged over the ventral and lateral views. We also describe three baselines and compare them with our approach.

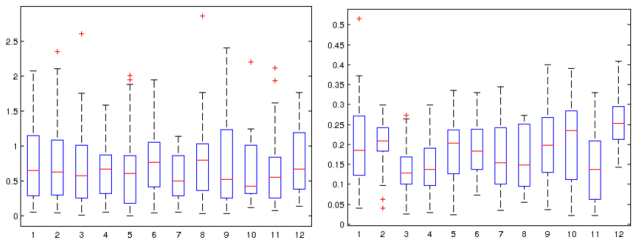
Table 1 shows our results of estimating the right number of the tooth types for each bat species when it is treated as "unknown". In our experiments, learning of the two models on the ventral and lateral views identified the same set and number of teeth. Therefore, Table 1 does not make a distinction between the two models. As can be seen, our model identification is highly accurate with at most 1 falsely included and 1 falsely excluded tooth type per species.

Next, we test dense image matching by evaluating the Euclidean distances between the transferred annotations and the ground truth in the unannotated images. These distances are normalized to the distance between the centers of the nasal aperture and the incisor tooth (I1), which are typically 8-10 pixels apart. Ideally, the normalized Euclidean distances should be zero. Fig. 3a shows the box plots of the normalized Euclidean distances averaged for every tooth across all eight bat species. The x-axis of the box plots enumerates 12 teeth belonging to teeth. On each box, the central red mark is the mean distance; the box edges correspond to 25% and 75% of the distances; the dashed lines extend to the most extreme data points not considered outliers; and outliers are plotted individually as red crosses. As can be seen, most means of the normalized Euclidean distances are greater than 0.5. A few outliers in the transferred annotation have the normalized distance greater than 2.5. These results demonstrate that the transferred annotations are noisy.

Our tooth localization results using the tree model are

Part Name	N	I1	C	P4	P5	M1	M2	# incorr
Artibeus	1	1	1	1	1	1	1	
	1	1	1	1	1	1	1	0
Desmodus	1	1	1	0	1	1	0	
	1	1	1	0	0	1	1	2
Glossophaga	1	1	1	1	1	1	1	
	1	1	1	0	1	1	1	1
Molossus	1	1	1	0	1	1	1	
	1	1	1	1	1	1	1	1
Mormoops	1	1	1	1	1	1	1	
	1	1	1	1	1	1	1	0
Noctilio	1	1	1	0	1	1	1	
	1	1	1	1	1	1	1	1
Saccopteryx	1	1	1	1	1	1	1	
	1	1	1	1	1	1	1	0
Trachops	1	1	1	1	1	1	1	
	1	1	1	1	1	1	1	0

Table 1: Results of our model identification using the t-test (Sec. 6). The 8 bat species are organized in rows. For each species, the top row indicates the ground truth presence "1" or absence "0" of the tooth, and the bottom row indicates our results. The teeth are: I1: mesial upper incisor; C: upper canine; P4: central upper premolar; P5: distal upper premolar; M1: mesial upper molar; M2: central upper molar. For the ventral views, we also account for a special part, N: nasal aperture.



(a) (b)

Figure 3: Box plots of (a) annotation transfer error, and (b) our part localization error measured as normalized Euclidean distances from the ground truth locations of every tooth averaged over the eight bat species. The x-axis enumerates 1: I1, 2: Nose, 3: C right, 4: P4 right, 5: P5 right, 6: M1 right, 7 M2 left, 8: C left, 9: P4 left, 10: P5 left, 11: M1 left, 12 M2 left. On each box, the central red mark is the mean distance, the edges of the box are 25% and 75% of the distances, the dashed lines extend to the most extreme errors not considered outliers, and outliers are plotted individually as red crosses.

shown in Fig. 3b. Localization of only correctly predicted tooth types is tested. As in the above image matching evaluation, localization error is measured as the normalized Euclidean distance between the centers of detected teeth and their ground truth in the unannotated images. Fig. 3b shows

Species	DPM	MED	Ours	SOSVM+GT
Artibeus	0.12	0.10	0.10	0.08
Desmodus	0.42	0.91	0.55	0.06
Glossophaga	0.33	0.21	0.19	0.11
Molossus	0.25	0.17	0.18	0.12
Mormoops	0.55	0.34	0.31	0.17
Noctilio	0.17	0.15	0.12	0.12
Saccopteryx	0.31	0.19	0.19	0.09
Trachops	0.16	0.18	0.09	0.06
Average	0.29	0.28	0.22	0.10

Table 2: Tooth localization error measured by normalized Euclidean distance from ground truth locations and averaged over all the teeth. DPM [17]: the mean of transferred annotations is used for initialization in training; MED: a variant of our approach that uses a mixture of Euclidean distances for the loss function in training; Ours: our approach; SOSVM+GT: SOSVM that uses ground truth part annotations for training.

the box plots of the average localization error for every tooth across all eight bat species. The mean localization errors are about 0.2 (*i.e.*, 4–5 pixels), which is roughly a 65% reduction compared to the errors in annotation transfer (Fig. 3a). This shows that our zero-shot learning greatly improves localization accuracy compared to the naive approach of just transferring the annotations without learning.

Finally, we compare our localization error with that of the following three baselines. The first baseline is the DPM presented in [17]. For initialization of latent locations of the DPM parts, we use the mean locations of the transferred annotations for each tooth. The second baseline, called MED, is a simpler variant of our approach that uses a mixture of Euclidean distances ($\eta = 0$), instead of our mixture of truncated Euclidean distances, for the loss function. The third baseline is the SOSVM trained directly on the ground truth annotations, denoted as SOSVM+GT. Table 2 compares our method against these three baselines. Note that in 7 of the 8 species, our method is at or near the best. We make serious errors only on *Desmodus*, a highly derived bat species adapted to blood feeding that has a very unusual set of teeth. For 4 of the species, our method is within 0.05 of SOSVM+GT, although unlike SOSVM+GT we do not have access to ground truth annotations. Interestingly, training with mean transferred annotations (MED) is slightly better than the DPM. This suggests the importance of employing a part-level loss function. But for some cases, MED obtains worse results. In particular, when image correspondences are poor, it tends to fail.

8. Conclusion

We present a new approach to zero-shot learning for detection and localization of object parts. This approach aimed to facilitate biological studies on tooth types and location in images of skulls from eight bat species. For the "unknown" bat species, we learn a tree model of the bat's teeth by transferring tooth annotations from images of "known" bat species to the unannotated images of the "unknown" bat species. The model is then applied to detect and localize the teeth in the unannotated images. Our experiments on 320 images of skulls of eight bat species demonstrate that we generally outperform baselines and often achieve performance close to an upper bound tree model trained on full ground truth annotations.

Acknowledgements

This research has been sponsored in part by NSF DEB 1208272 to T. G. Dietterich, NSF DEB 1208270 to M. A. O'Leary and NSF DEB 1208306 to N. B. Simmons.

References

[1] A. Farhadi, I. Endres, and D. Hoiem. Attribute-centric recognition for cross-category generalization. *CVPR*, 2010. 2

[2] P. Feltzenswalb and D. Hutenlocher. Pictorial structures for object recognition. *IJCV*, 61(1):55–79, 2005. 3

[3] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *PAMI*, 2010. 3, 4

[4] C. H. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. *CVPR*, 2009. 2

[5] P. O. Lewis. A likelihood approach to estimating phylogeny from discrete morphological character data. *Syst. Biol.*, 50(6):913–925, 2001. 1

[6] J. J. Lim, R. Salakhutdinov, and A. Torralba. Transfer learning by borrowing examples for multiclass object detection. *NIPS*, 2011. 2

[7] C. Liu, J. Yuen, and A. Torralba. Nonparametric scene parsing: label transfer via dense scene alignment. *CVPR*, 2009. 3, 4

[8] M. A. O'Leary, J. I. Bloch, J. J. Flynn, T. J. Gaudin, A. Giallombardo, N. P. Giannini, S. L. Goldberg, B. P. Kraatz, Z.-X. Luo, J. Meng, X. Ni, M. J. Novacek, F. A. Perini, Z. Randall, G. W. Rougier, E. J. Sargis, M. T. Silcox, N. B. Simmons, M. Spaulding, P. M. Velazco, M. Weksler, J. R. Wible, and A. L. Ciranello. Response to comment on "the placental mammal ancestor and the post-k-pg radiation of placentals". *Science*, 341:613, 2013. 1

[9] M. A. O'Leary and S. Kaufman. Morphobank: phylophenomics in the "cloud". *Cladistics*, 27:1–9, 2011. 1

[10] N. Ratliff, J. A. Bagnell, and M. Zinkevich. Subgradient methods for structured prediction. *AISTATS*, 2007. 4

[11] R. Salakhutdinov, J. Tenenbaum, and A. Torralba. Learning to learn with compound hd models. *NIPS*, 2011. 2

[12] A. Torralba, K. P. Murphy, and W. T. Freeman. Sharing features: efficient boosting procedures for multiclass object detection. *CVPR*, 2004. 2

[13] I. Tsochantaridis, T. Hofmann, T. Joachims, and Y. Altun. Support vector machine learning for interdependent and structured ouput spaces. *ICML*, 2004. 3, 4

[14] Y. Wang and G. Mori. A discriminative latent model of object classes and attributes. *ECCV*, 2010. 2

[15] J. Wiens. The role of morphological data in phylogeny reconstruction. *Syst Biol*, 53:653–661, 1999. 1

[16] E. O. Wilson. The encyclopedia of life. *Trends in Ecology and Evolution*, 18:77–80, 2003. 1

[17] Y. Yang and D. Ramanan. Articulated pose estimation with flexible mixtures-of-parts. *CVPR*, 2011. 4, 6

[18] L. Zhu, Y. Chen, A. Torralba, W. Freeman, and A. Yuille. Part and appearance sharing: Recursive compositional models for multi-view multi-object detection. *CVPR*, 2010. 2