Attribute-Based Detection of Unfamiliar Classes with Humans in the Loop

Catherine Wah Serge Belongie
Department of Computer Science and Engineering
University of California, San Diego
{cwah, sjb}@cs.ucsd.edu

Abstract

Recent work in computer vision has addressed zero-shot learning *or* unseen class detection*, which involves categorizing objects without observing any training examples. However, these problems assume that attributes or defining characteristics of these unobserved classes are known, leveraging this information at test time to detect an unseen class. We address the more realistic problem of detecting categories that do not appear in the dataset in any form. We denote such a category as an* unfamiliar class*; it is neither observed at train time, nor do we possess any knowledge regarding its relationships to attributes. This problem is one that has received limited attention within the computer vision community. In this work, we propose a novel approach to the unfamiliar class detection task that builds on attribute-based classification methods, and we empirically demonstrate how classification accuracy is impacted by attribute noise and dataset "difficulty," as quantified by the separation of classes in the attribute space. We also present a method for incorporating human users to overcome deficiencies in attribute detection. We demonstrate results superior to existing methods on the challenging CUB-200-2011 dataset.*

1. Introduction

A recent trend in the computer vision community is the use of high-level visual features or attributes as semantic cues in addressing various problems. Attributes can be used to describe and differentiate classes [18, 12, 16], and information about attributes and their values can be exploited in order to perform *unseen class detection*, where the goal is to categorize objects into classes for which we have no training examples (*i.e.* the train and test sets are disjoint).

For example, a visual recognition system for North American bird species that has not been trained on images of Indigo Buntings can still maintain a database listing the species' distinguishing attributes, such as having blue bellies and black legs. Should the system then detect the presence of these attributes in a test example, it can predict with high confidence that the bird in question is an Indigo Bunting without ever having seen examples of that species.

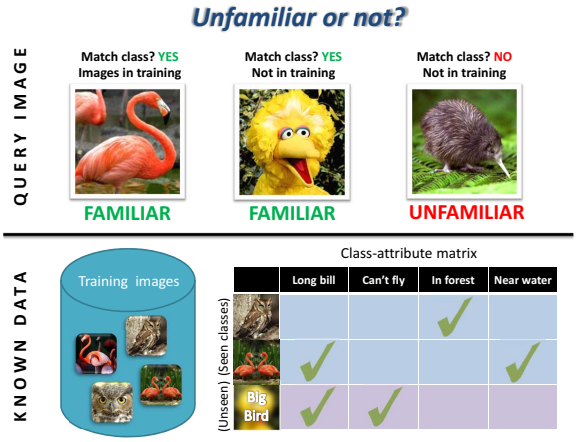


Figure 1: We address *unfamiliar class detection*, in which the goal is to predict if an input image belongs to an unfamiliar class (right query), defined by its lack of observed training examples and unknown class-attribute relationships. We contrast this with *unseen classes* (middle query), which do not occur in training but have entries in the class-attribute matrix. *Seen classes* (left query) will occur both in training as well as the class-attribute matrix. Unfamiliar class detection is a challenging problem, given that the system must be able to distinguish between a difficult example of a known class and a truly unfamiliar class.

These applications typically assume that sufficient knowledge of the characteristics of unseen classes is provided or can be extracted prior to test time [18, 12, 16, 11, 25, 31]. In this work, we study the related and more challenging problem of detecting *unfamiliar classes*. The distinction between an unfamiliar class and an unseen class lies in the amount of knowledge regarding the category that is available to the system. While examples belonging to unfamiliar and unseen classes are both unobserved during

training, unfamiliar classes lack entries in the matrix defining class-attribute associations; in contrast, an unseen class' relationship to attributes is known. In the North American bird recognition system example, this would be akin to submitting an image of a Kiwi (a bird species native to New Zealand) and asking the system to recognize it. The system has no prior knowledge about what attributes describe a Kiwi and has never seen an image of one before (see right query image in Figure 1).

The problem of unfamiliar class detection marks a departure from the predominant closed-set approach to visual categorization, in which the goal is to classify test examples as one of a fixed set of possible classes. While this closed-set approach has enabled scientific progress within the field, the datasets used may suffer from biases and restrictions that can be exploited by certain algorithms [30, 35].

This focus on closed-set problems is in part because they are more feasible to address. Standard recognition algorithms have been trained on datasets [13, 14, 32, 10] of up to hundreds of basic-level visual categories; however, the set of human-recognizable basic-level categories has been estimated to be roughly $30,000$ [4], and these datasets cover only a small portion of this variety. Large-scale categorization datasets such as ImageNet [7] provide significantly more coverage of the visual category space; nevertheless, it is in reality very challenging, if not impossible, to inventory all visual objects with category labels, and there will always exist classes that are unfamiliar to any given dataset.

As such, detecting unfamiliar classes becomes a significant problem as recognition systems improve and are deployed in the wild. This issue is especially relevant for fine-grained categorization systems [24, 17, 23, 6, 37], as these systems are often of limited scope (*e.g.* North American birds, plants of the Northeast U.S.). Moreover, human users will not necessarily have the domain knowledge to perform verification or identify if the true class is in the dataset, as the distinctions between *fine-grained categories* (*e.g.* Mallard, Cardinal), which comprise a *basic-level category* (*e.g.* bird), can be subtle.

While unseen class detection is similar in spirit, unfamiliar class detection is a more salient problem in practice, yet also one that has been widely overlooked in the computer vision community. Our goals in this paper are to study this problem by means of an attribute-based approach and analyze the specific challenges associated with it. Attributes are a powerful high-level representation of visual features, capable of describing an exponential number of classes. Furthermore, they are often semantically meaningful to human users. In a problem as challenging as detecting unfamiliar classes, it is important to have a means of incorporating humans into the loop, as human users can be engaged at test time to bring performance up to a desired level.

However, the use of attributes contributes to various challenges in unfamiliar class detection. First, there is an indeterminate amount of error in attribute values that can arise from either attribute detectors or user labels. Second, the number of differing attributes between classes quantifies how distinct the classes are; if the classes have low separation in the attribute space, the dataset will be more "difficult" and less robust to attribute noise.

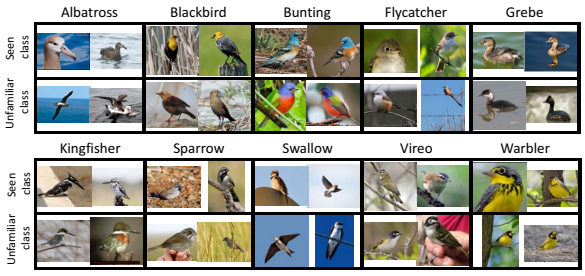


Figure 2: **Bird-Families Dataset.** From *CUB-200-2011* [38], we created a 40-class dataset of 20 manually identified taxonomic families. One class is kept as a seen class and the other class is deemed unfamiliar. Refer to Figure 4 for additional pairs.

Our contribution is three-fold: (1) we present a novel attribute-based algorithm for unfamiliar class detection; (2) we empirically demonstrate how accuracy in detecting unfamiliar classes is affected by the aforementioned challenges; and (3) we support the addition of humans into the loop in the form of attribute responses. We demonstrate results on the *CUB-200-2011* [38] dataset.

The paper is organized as follows. In Section 2, we discuss related work. In Section 3, we formalize the unfamiliar class detection problem and describe our approach. We review implementation details in Section 4, and in Section 5, we present our experiments and discuss our results.

2. Related Work

Recent work addressing the problem of zero-shot learning or unseen class detection has taken advantage of the generality of high-level attribute descriptions [18, 12, 16, 11, 25]. Some methods treat zero-shot learning as a nearest-neighbor problem, classifying a test image as the category with the most similar attribute description [12, 25]. These applications typically assume that sufficient knowledge of the characteristics of the unseen classes is provided or can be extracted prior to test time. These characteristics can come in various forms, including binary attribute descriptions [25] or relative relationships between attributes [27]. Of the limited works that do address unfamiliar class detection [12, 19], the focus is on basic-level categories, whereas we focus on detecting unfamiliar fine-grained categories.

Approaches to large-scale multi-class classification ad-

dress the computational infeasibility of testing against all possible classes by using a hierarchical tree structure of classifiers [15, 1, 8]. However, these methods do not deal with unfamiliar classes that may be encountered at test time. Other work directly addresses the novel category scenario in large-scale recognition by having a user provide a set of images belonging to an unfamiliar class [3]; the focus in this case is on retrieval, as the novelty of the test class is known.

For unfamiliar classes, the primary challenge is that the class-attribute relationships are unknown. Determining class-attribute associations is a non-trivial task. The associations can potentially be manually assigned by human subjects or experts with appropriate domain knowledge [18, 6]; other work has focused on determining these associations either automatically by mining natural language resources [31, 2] or interactively by using human guidance [26]. While the attributes can be harvested with varying levels of automation, the class-attribute descriptions are generally assumed to be determined or obtained prior to testing. In this work, our goal is to detect unfamiliarity, and we do not aim to determine true class-attribute descriptions.

Mahajan et al. [20] address a problem similar to unfamiliar class detection but make some different assumptions. Their method learns the class-attribute relationships in addition to the attribute classifiers, eliminating the need for prior knowledge of these associations. However, the number of unfamiliar classes must be specified beforehand. We do not place such restrictions on the unfamiliar classes; examples from both seen and unfamiliar classes occur at test time, and we assume that all classes (seen and unfamiliar) can be characterized by the same super set of attributes.

In the machine learning literature, there is significant relevant work on novelty detection [21, 22], which includes techniques such as one-class classification methods [34, 5]. We note that among fine-grained visual categories of the same basic-level category, the interclass variation lies at the attribute level, and an unfamiliar class represents an unknown combination of these attributes. A one-class SVM cannot adequately characterize these fine-grained distinctions, which we empirically demonstrate in Section 5.1.

Others have addressed the open-set problem in recognition, where classes not seen in training may occur in testing; this issue is frequently encountered in the biometric verification setting, such as for face recognition or matching [33]. This work presents an approach to unfamiliar class detection in the visual categorization setting that uses high-level attribute features and is not application specific.

3. Approach

In this section, we introduce our algorithm for performing unfamiliar class detection using knowledge of attributes and present a method for incorporating user responses.

3.1. Problem formulation

Given an image x, our goal is to predict whether the true object class c for x belongs to the set of seen classes $c \in \mathcal{S}$, where $|\mathcal{S}| = C$, or is an unfamiliar class $c \in \mathcal{U}$, where $\mathcal{U}$ is the set of unfamiliar classes. The random variable $g \in \{0, 1\}$ denotes this characteristic of being unfamiliar. For fine-grained categories, we assume that images from seen and unfamiliar classes all contain an object belonging to the same basic-level category, that is, we do not address detecting object presence.

We assume that all classes $c \in \mathcal{S} \cup \mathcal{U}$ can be represented using a shared vocabulary of A attributes (*e.g.* striped wing, red crown, etc.) and can be described with a unique deterministic vector of attributes $\mathbf{a}^c = [a_1^c, \ldots, a_A^c], a_i^c \in 0, 1$, as in [18]. Unfamiliar classes differ from seen classes in that their attribute values and distributions are not known.

This assumption is reasonable given a sufficiently large attribute vocabulary, with which we can theoretically represent an exponential number of classes; in practice, it enables us to capture a significant amount of variation in attributes that may represent new categories. Detecting an unfamiliar class therefore involves predicting an unfamiliar combination of attributes, such that it does not correspond with any known class-attribute relationships.

We note that if an unfamiliar class is detected, a necessary task is to verify this prediction. In addition to verifying it is indeed a new class, one potentially would want to assign a descriptive label to this new class, as well as determine its true attribute description. Both the verification and naming of unfamiliar classes are non-trivial tasks, and we assume that they will be performed by an expert with necessary domain knowledge. This expert interaction stage would occur offline and is a different line of work. We do note that providing an estimate for unfamiliarity is still of value in this scenario, as it reduces reliance on the expert.

3.2. Detecting unfamiliar classes

Our goal in unfamiliar class detection is to estimate the probability of an incoming test example belonging to an unfamiliar class in $\mathcal{U}$. We first define the per-class probabilities, which are used in determining unfamiliarity. The probability of the object class c belonging to any of the individual seen classes in $\mathcal{S}$, given the image pixels x, is $p(c|x)$ and can be determined in terms of attributes:

$$p(c|x) = \int_{\mathbf{a}} p(c|\mathbf{a}, x) p(\mathbf{a}|x) dx. \quad (1)$$

We assume a direct-class attribute model, such that $p(c|\mathbf{a}, x)$ is nonzero only when $\mathbf{a} = \mathbf{a}^c$, where $\mathbf{a}^c$ is the ground truth attribute vector for class c. Together with our assumption that the class can be fully determined by a unique attribute membership vector, the integral of Equa-

tion 1 can then be expressed as:

$$p(c|x) = p(c|\mathbf{a}^c, x)p(\mathbf{a}^c|x) = p(c|\mathbf{a}^c)p(\mathbf{a}^c|x). \quad (2)$$

If $\forall c' \neq c$ then $\mathbf{a}^{c'} \neq \mathbf{a}^c$, this can be represented as $p(\mathbf{a}^c|x)$.

Estimating the probability of object class $c \in \mathcal{U}$ is equivalent to estimating how unlikely a test example belongs to any seen class in $\mathcal{S}$:

$$p(g|x) = \prod_c (1 - Zp(c|x)), \quad (3)$$

where $p(g|x)$ is the probability the class is unfamiliar, and Z represents a normalization term $\beta + \alpha \sum_{c=1}^{C} p(c|x)$. We cannot assume the true class of a test example belongs to the set of seen classes in $\mathcal{S}$ and normalize directly with $\prod_c p(c|x)$, so we learn α and β terms on a validation set to maximize the log-likelihood $\sum_j \log p(g_j|x_j)$.

The validation set contains seen as well as unfamiliar examples. Note that the unfamiliar examples in validation and testing are drawn from disjoint sets of categories. While we are unable to learn parameters that optimize for unfamiliar classes in the test set, we assume that all possible unfamiliar classes are drawn from the same basic-level category and thus validate on other classes within this basic-level category. Intuitively, classes that share a basic-level category are likely to have visual similarities (*e.g.* birds share body parts like beaks and wings).

3.3. Incorporating computer vision

Under the assumption of mutual independence of attributes, we estimate the attribute probabilities using binary attribute classifiers. We convert the attribute classification scores $z_i = a_i^c \langle \mathbf{w}_i, x \rangle$, $i \in 1 \dots A$, to probabilities by fitting a sigmoid function $\sigma(\gamma_a z_i)$ [29]:

$$p(\mathbf{a}^c|x) = \prod_{i=1}^{A} p(a_i^c|x) = \prod_{i=1}^{A} \sigma(\gamma_a z_i), \quad (4)$$

where $a_i^c \in \{-1, 1\}$ and z_i is a linear function of the image pixels. The parameters $\mathbf{w}_i$ are the learned weights of each of the attribute classifiers. The sigmoid parameter γ_a is learned on a validation set (as in Section 3.2) for each attribute a_i, to maximize the log-likelihood $\sum_j \log p(a_i|x_j)$. In practice, this approach worked better than maximizing over the log-likelihood probability of belonging to an unfamiliar class, as attribute classifier performance and robustness play a large role in accurately predicting unfamiliarity.

3.4. Incorporating user responses

An advantage of an attribute-based approach is the ability to incorporate human responses. Our framework for unfamiliar class detection supports the addition of human users, who can boost performance by replacing outputs of poor attribute detectors. We model the set of user attribute responses U in a similar fashion as [37, 6]. Given the assumptions from Section 3.2, we express the per-class probabilities $p(c|U, x)$ as:

$$p(c|U, x) = \frac{p(U|\mathbf{a}^c, x)p(\mathbf{a}^c|x)}{p(U|x)}. \quad (5)$$

A user's perception of an attribute value a_i is denoted as a random variable $\tilde{a}_i$, and we assume attribute values are perceived independently:

$$p(U|x) = \prod_{\tilde{a}_i \in U} p(\tilde{a}_i). \quad (6)$$

It remains for us to estimate $p(U|\mathbf{a}^c, x)$. We assume that a user's perception of attribute $\tilde{a}_i$ is dependent only on the ground truth attribute a_i^c:

$$p(U|\mathbf{a}^c, x) = \prod_{\tilde{a}_i \in U} p(\tilde{a}_i|a_i^c)^\gamma = \exp\{\sum_{\tilde{a}_i \in U} \gamma \log p(\tilde{a}_i|a_i^c)\}, \quad (7)$$

and as in Section 3.2, we learn the parameter γ with cross-validation. Using parameters learned on Mechanical Turk responses, we can estimate $p(\tilde{a}_i|a_i^c)$ for each attribute.

4. Implementation Details

In this section, we briefly discuss the datasets, features, and computer vision algorithms used.

4.1. Datasets

We created two collections drawn from *CUB* for experiments: *All-Birds*, and *Bird-Families*. The *All-Birds* collection uses 150 of the 200 available categories; 100 of those classes were randomly selected to be kept as seen classes, and another 50 randomly selected classes were denoted unfamiliar. Images from the unfamiliar categories were then removed from the train set. We also removed corresponding entries in the class-attribute matrix for the unfamiliar classes. At test time, we included examples from the 150 classes. We performed 5-fold cross validation on random sets of 10 unfamiliar classes drawn from the remaining 50 classes that did not appear in training or testing.

The *Bird-Families* collection consists of classes selected from 20 manually identified taxonomic families within the *CUB* dataset (see Figures 2 and 4 for examples). We determined families based on taxonomic classification (e.g. Least Terns and Arctic Terns are considered to be from the family Tern), keeping the 20 families with the most classes. We note that species that are considered similar in terms of their scientific classification often share many visual features, but visual similarity is not a precondition.

From each family, we randomly selected one class to be included in the dataset as seen, and we selected a second class as an unfamiliar category. The resulting collection eliminates many classes that fall under the same family

(for instance, *CUB* includes 25 species of Warblers) and are commonly confused. At the same time, it remains a challenging dataset: due to the selection of family pairs, unfamiliar classes may closely resemble seen classes. The validation set consists of 5 classes drawn randomly from the remaining 160 classes in *CUB*. For both bird datasets, we used roughly 30 images per class to train our attribute classifiers, and 15 per class to test. We obtained the ground truth class-attribute matrix by binarizing class-averaged user responses.

4.2. Features and learning

For the two collections, we used localized features for all attributes. Each attribute is associated with a certain part of the bird; object-level attributes such as Has Primary Color Blue, Has Size Large, and Has Shape Perching-Like are associated with the entire bounding box of the bird. Similar to [6], we extracted features using [36] that include vector-quantized color histograms, geometric blur and color/gray SIFT features using spatial pyramids, from patches around ground truth part locations.

We trained binary classifiers for each attribute on the concatenated feature histograms with linear SVMs. In general, we expect more sophisticated learning algorithms to yield attribute classifiers of greater discriminative power that can boost classification accuracy as well as unfamiliar class detection accuracy. We discuss the effects of attribute classifier performance in Section 5.1.

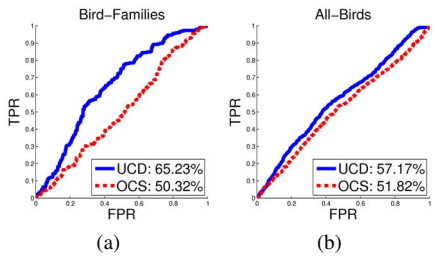
5. Experiments

In this section, we present our experimental results. We observe empirically in Sections 5.2, 5.3, and 5.4 the roles that attribute noise, class attribute variation, and users respectively play in detecting unfamiliar classes accurately.

5.1. Unfamiliar class detection

We present results using our method on both datasets in Table 1 and Figure 3. We measure performance in unfamiliar class prediction as the area under the ROC curve (AUC); the advantage of using this metric is that it is invariant to priors. On our *Bird-Families* dataset, we obtain an AUC of 0.652 for predicting unfamiliar classes. We note the only similar experiment that we are aware of was presented by Farhadi et al. [12] for rejecting unknown categories. They reported an AUC of 0.6 for 20 Pascal object classes using 65 attributes, with significantly more training data and focusing on basic-level categories.

We compare our performance to a one-class SVM [34], which attempts to find a smooth boundary enclosing a region, while separating out a fixed fraction ν of the training data. Using the feature vectors described in Section 4.2 with an RBF kernel, the one-class SVM performed only marginally better than random chance. Given the inherent low interclass variation of fine-grained categories, this method is unable to find a suitable hyperplane to separate the seen from the unfamiliar.



(a) (b)

Figure 3: ROC curves for unfamiliar class detection with the *Bird-Families* (3a) and *All-Birds* (3b) datasets.

Dataset	*#SC*	*#UC*	*Clf.*	*UCD*	*OCS*	*TCS*
Families	20	20	0.708	0.652	0.503	0.512
All-Birds	100	50	0.284	0.572	0.518	0.514

Table 1: We present results for the multi-class classification of seen classes (average accuracy on seen class test images), and report AUCs for unfamiliar class detection (UCD); a one-class SVM (OCS) with $\nu = 0.5$; and a two-class SVM (see Section 5.1 for more details). Also shown are the numbers of seen (SC) and unfamiliar classes (UC).

A second baseline we compare to is a two-class SVM. We train a binary classifier to recognize unfamiliar classes, in which negative examples are drawn from all familiar classes, and positive examples are drawn from a held-out set of unfamiliar categories consisting of 5 classes selected at random for *Bird-Families* and 25 for *All-Birds*. Performance averaged over 10 folds of cross validation is comparable to the one-class SVM.

We also report accuracy for the multi-class classification task on the seen classes to demonstrate the difficulty of these datasets without even considering the unfamiliar class problem. Per-class probabilities for the test examples belonging to seen classes are computed using the model in [18].

For both multi-class classification and unfamiliar class detection, the discriminativeness of the individual attribute classifiers clearly affects performance. The average AUC for the classifiers is 70.94%; low performance may be due to factors such as imprecise localization. In general, we note that better attribute classifiers are likely to produce higher overall accuracy for the classification task and consequently, unfamiliar class detection, as the prediction of unfamiliarity is based on the per-class probabilities.

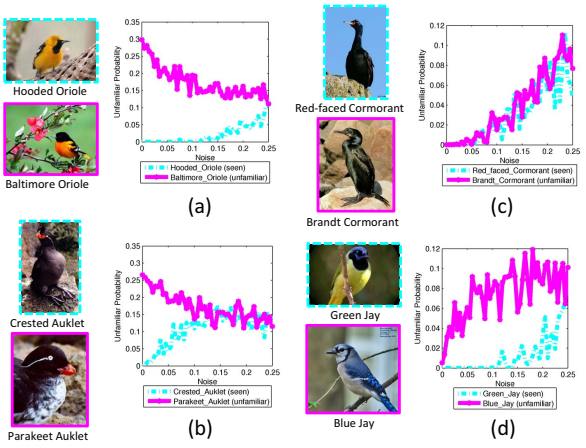


Figure 4: **Modeling attribute noise.** For several bird family pairs 4(a-d), we observe how unfamiliar class probabilities change as noise is added to the attributes. Noise is quantified as the probability an individual attribute value is incorrect. Each family is represented by one seen (cyan dashed lines) and one unfamiliar class (solid magenta lines). The curves represent the probability for a certain class averaged over all test images (Section 5.2).

5.2. Modeling attribute noise

As noted in Section 5.1, the accuracy of the classifiers themselves play a large role in the overall success in predicting unfamiliarity. An attribute classifier's performance can be characterized in terms of noisiness if the output is treated as a binary response—a poorly performing attribute classifier will therefore detect an attribute with a high rate of error or noise. Similarly, user-labeled attributes exhibit ambiguity and errors due to differences in perception.

The amount of noise present in detecting attributes, whether it arises from human users or attribute classifiers, impacts how well one can detect unfamiliar classes. In this experiment, we observe the effect of attribute noise in unfamiliar class detection, focusing on the *Bird-Families* dataset as it allows us to examine this effect for similar and commonly confused classes.

We train a naïve Bayes classifier on the image-level attribute labels, using a Beta prior to improve robustness and estimating parameters with maximum a posteriori estimation. Because we wish to observe the effect of varying amounts of noise at test time, we use ground truth class-attribute values and simulate the addition of noise, which is added independently and uniformly to all attributes.

The added noise represents the probability that an attribute bit value will be flipped. For example, assume the attributes are determined at test time with 5% noise. In a 200-attribute vocabulary, roughly 10 attributes will then be incorrectly determined, suggesting that the classes must differ from each other by at least 10 attributes, in order to be robust to noise. Error-correcting output codes (ECOCs) quantify error in a similar manner [9].

We focus on pairs of related classes within the same family. For each family, recall that one class appears in the dataset as a seen class, and the other is considered an unfamiliar class. As noise is added to the attribute values, we observe the probability that test examples from the two classes are unfamiliar (see Figure 4). In order to better understand what is happening at the class level, the probabilities for test images are averaged over the true class.

Referring to Figures 4(a-b), we observe that with no added noise, the unfamiliar classes are predicted with high probability of being unfamiliar, as compared to the corresponding seen classes. We do not take into account any bias or priors, so a threshold on the unfamiliar probability is not determined. Regardless, it is important to note the clear separation between how likely an unfamiliar class is indeed unfamiliar versus how likely a seen class is considered unfamiliar. As noise is added and errors are introduced to the attribute values, the probability of being unfamiliar for both classes in a family tends to converge. This indicates a confusion between seen and unfamiliar classes; after exceeding a certain level of noisiness, it is not possible to discern between them as both are equally likely to be unfamiliar.

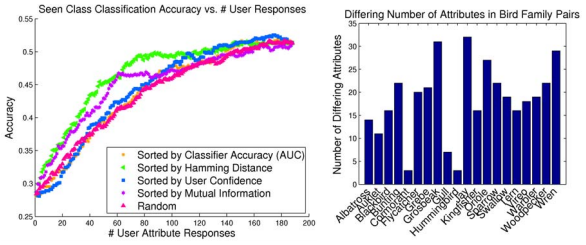
5.3. Modeling class attribute variation

Another dimension to unfamiliar class detection deals with the amount of class attribute variation, which can be thought of as the "difficulty" of the dataset. We quantify it as the average number of attributes that differ between classes; this metric is computed as the average Hamming distance between all pairs of binary class attribute vectors.

If the classes are well separated in the multidimensional attribute space, then the difficulty is considered low; on the other hand, if the classes in the dataset share many attributes, it is more challenging to distinguish between them, especially with noisy attribute detection, and the dataset is considered difficult. As noted in the previous section, the minimum Hamming distance between any pair of classes in the dataset quantifies the limiting factor in terms of robustness to noise, similar to the quality of an ECOC [9].

In Figure 4(c), we note that even in the absence of added noise, Brandt Cormorants are not correctly detected as unfamiliar. The unfamiliar Brandt Cormorant test examples are instead mistaken for their seen class counterpart, the Red-faced Cormorant. The pair of classes is visually very similar, with the categories differing only by 3 attributes (see Figure 5b).

While the family pairs represent taxonomically similar species, the classes are not necessarily visually similar and may share more attributes in common with other classes.



(a) (b)

Figure 5: 5a: Seen class classification observed as we query humans for attribute values. 5b: The Hamming distance between classes of the same family from *Bird-Families*.

For Figure 4(d), we note that the Blue Jay and the Green Jay differ by the most visual attributes (32) out of all family pairs. As such, Blue Jays are not mistaken for Green Jays; however, they are also not likely to belong to an unfamiliar class. The Blue Jays are in fact getting confused with Lazuli Buntings, from which they differ by fewer attributes (13).

5.4. Modeling users

To overcome noise in attribute classifiers, which is compounded by the inherent difficulty of the dataset at hand, we can incorporate human users into the pipeline. In order to observe the role users play in unfamiliar class detection, we perform an experiment in which we select attribute questions one at a time to ask users.

We investigate several different methods for selecting what attribute value to query (Figure 6), using the *All-Birds* dataset: (1) based on the performance of the trained attribute classifiers, as ranked by the AUC; (2) based on how well an attribute can discriminate the classes in the dataset; (3) based on observed user confidence in training; (4) based on mutual information (MI); and (5) at random. We determine how well an attribute can discriminate the classes by sorting the attributes based on their average Hamming distance to all other attributes. User confidence on a per-attribute basis is observed on the training set, and we query users of attributes that they tend to answer with the highest certainty (users are given the options *definitely*, *probably*, and *guessing* when providing attribute values). The fourth method ranks the attributes based on mutual information, such that attributes that share less information with other attributes are selected first.

In Figure 6, we observe that we are able to improve unfamiliar class detection accuracy by querying users selectively, suggesting that despite variance in user responses, we are able to leverage them to overcome poor attribute detections. We note that the order in which users are queried has an impact on unfamiliarity detection performance. For example, the first two attributes queried using the Hamming distance and MI methods are Has Eye Color Black and Has Belly Pattern Solid, which are two of the most common attributes found in the seen classes. These attributes are useful in detecting unfamiliar classes, because when they are not detected in a test example, then the example is less likely to belong to any of the seen classes. The attributes queried later tend to occur infrequently in the set of seen classes and thus are less informative in general.

We observe that by using user responses that are answered with high certainty, without considering consistency in response, we can boost unfamiliar class detection performance significantly. While user consistency is not explicitly observed, it still may contribute to improved unfamiliarity detection performance early on. As more questions are answered, the system incorporates less reliable attribute values, and this increased noise in attribute values (Figure 4) may cause the drop in the curve after 60 user responses. At the same time, user certainty has no observable impact on seen class classification performance (Figure 5a), suggesting that the issues of detecting unfamiliarities and classification require different approaches and should be addressed individually.

6. Conclusion

In this work, we have presented an attribute-based framework for unfamiliar class detection that supports the use of humans in the loop, empirically observing the roles that attribute noise and variation play in the task of unfamiliar class detection. Success at detecting unfamiliar classes is influenced by how distinct the classes in the dataset are to each other, as there inevitably will be noise in the attribute detection. Achieving better performance on unfamiliar class detection also necessitates improving accuracy in the classification of attributes and classes; however, we emphasize that these problems should be addressed separately. In order to improve attribute-based classification, one could take into account dependencies between attributes or consider multiple modalities in class attribute descriptions; these are directions for future work.

The emerging trend of computer vision entering the wild in the form of semi-automated field guides (Visipedia [28], Leafsnap [17], butterflies [39], etc.) indicates that we have only seen the tip of the iceberg for the unfamiliar class problem. This work offers progress towards a solution, and we hope to spur further discussion and study of this problem within the computer vision community.

7. Acknowledgments

The authors thank Steve Branson and Peter Welinder for helpful discussions and feedback. This work is supported by the Amazon AWS in Education program and the National Science Foundation Graduate Research Fellowship under Grant No. DGE0707423.

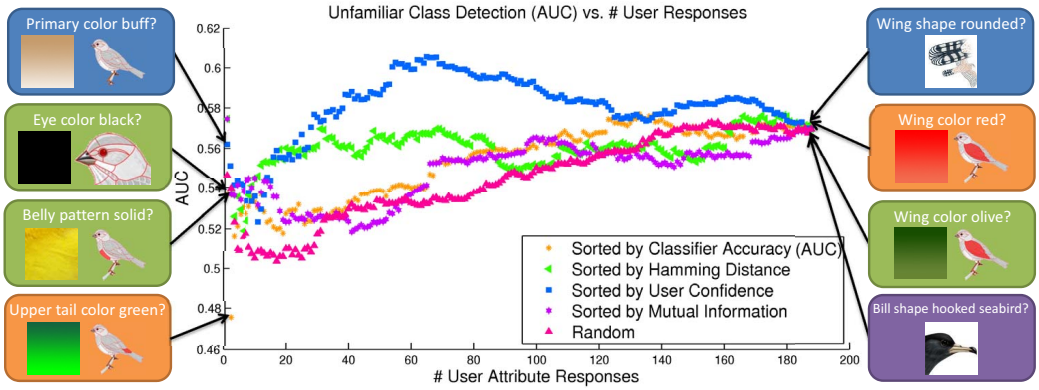


Figure 6: **Using humans in the loop.** We compare several methods of sequentially querying users and observe how each affects unfamiliar class detection performance. The methods are plotted according to the number of user responses so far. We show examples of the first and last attributes added using these methods. See Section 5.4 for more details.

References

[1] S. Bengio, J. Weston, and D. Grangier. Label embedding trees for large multi-class tasks. In *NIPS*, 2010.

[2] T. Berg, A. Berg, and J. Shih. Automatic attribute discovery and characterization from noisy web data. In *ECCV*, 2010.

[3] A. Bergamo et al. Picodes: Learning a compact code for novel-category recognition. In *NIPS*, 2011.

[4] I. Biederman. Recognition by components: a theory of human image interpretation. *Pyschological review*, 94:115–147, 1987.

[5] P. Bodesheim et al. Divergence-based one-class classification using gaussian processes. In *BMVC*, 2012.

[6] S. Branson et al. Visual recognition with humans in the loop. In *ECCV*, 2010.

[7] J. Deng et al. ImageNet: A large-scale hierarchical image database. In *CVPR*, 2009.

[8] J. Deng et al. Fast and balanced : Efficient label tree learning for large scale object recognition. In *NIPS*, 2011.

[9] T. Dietterich and G. Bakiri. Solving multiclass learning problems via error-correcting output codes. *Journal of Artificial Intelligence Research*, 2:263–286, 1995.

[10] M. Everingham et al. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://pascallin.ecs.soton.ac.uk/challenges/VOC/voc2012/index.html.

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

[12] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In *CVPR*, 2009.

[13] L. Fei-Fei et al. Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories. In *IEEE CVPR Workshops*, 2004.

[14] G. Griffin, A. Holub, and P. Perona. Caltech-256 object category dataset. Technical Report CNS-TR-2007-001, California Institute of Technology, 2007.

[15] G. Griffin and P. Perona. Learning and using taxonomies for fast visual categorization. In *CVPR*, 2008.

[16] N. Kumar et al. Attribute and simile classifiers for face verification. In *ICCV*, 2009.

[17] N. Kumar et al. Leafsnap: A computer vision system for automatic plant species identification. In *The 12th European Conference on Computer Vision (ECCV)*, October 2012.

[18] C. H. Lampert et al. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*, 2009.

[19] Y. J. Lee and K. Grauman. Object-graphs for context-aware category discovery. In *CVPR*, 2010.

[20] D. Mahajan et al. A joint learning framework for attribute models and object descriptions. In *CVPR*, 2011.

[21] M. Markou and S. Singh. Novelty detection: a review-part 1: statistical approaches. *Signal Processing*, 83:2481–2497, 2003.

[22] M. Markou and S. Singh. Novelty detection: a review-part 2: neural network based approaches. *Signal Processing*, 83:2499–2521, 2003.

[23] G. Martinez-Muñoz et al. Dictionary-free categorization of very similar objects via stacked evidence trees. In *CVPR*, 2009.

[24] M. Nilsback and A. Zisserman. Automated flower classification over a large number of classes. In *ICCVGIP*, 2008.

[25] M. Palatucci et al. Zero-Shot Learning with Semantic Output Codes. In *NIPS*, 2009.

[26] D. Parikh and K. Grauman. Interactively building a discriminative vocabulary of nameable attributes. In *CVPR*, 2011.

[27] D. Parikh and K. Grauman. Relative attributes. In *ICCV*, 2011.

[28] P. Perona. Vision of a visipedia. *Proceedings of the IEEE*, 98(8):1526 –1534, aug. 2010.

[29] J. Platt. Probabilities for SV machines. In *NIPS*, pages 61–74, 1999.

[30] J. Ponce et al. Dataset issues in object recognition. In *Towards Category-Level Object Recognition*, pages 29–48. Springer, 2006.

[31] M. Rohrbach et al. What helps where – and why? semantic relatedness for knowledge transfer. In *CVPR*, 2010.

[32] B. C. Russell et al. LabelMe: A database and web-based tool for image annotation. *IJCV*, 77(1-3):157–173, 2008.

[33] W. J. Scheirer, A. Bendale, and T. E. Boult. Predicting biometric facial recognition failure with similarity surfaces and support vector machines. In *IEEE CVPR Workshops*, 2008.

[34] B. Schölkopf, J. Platt, J. Shawe-Taylor, A. Smola, and R. Williamson. Estimating the support of a high-dimension distribution. *Neural Computation*, 13(7):1442–1471, 2001.

[35] A. Torralba and A. A. Efros. Unbiased look at dataset bias. In *CVPR*, 2011.

[36] A. Vedaldi, V. Gulshan, M. Varma, and A. Zisserman. Multiple kernels for object detection. In *ICCV*, 2009.

[37] C. Wah et al. Multiclass recognition and part localization with humans in the loop. In *ICCV*, 2011.

[38] C. Wah et al. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, Caltech, 2011.

[39] J. Wang et al. Learning models for object recognition from natural language descriptions. In *BMVC*, 2009.