The Animal ID Problem: Continual Curation

C. V. Stewart
Rensselaer Polytechnic Institute
stewart@rpi.edu

J. R. Parham, J. Holmberg
Wild Me
{parham,jason}@wildme.org

T. Y. Berger-Wolf
Ohio State University
berger-wolf.1@osu.edu

Abstract

Hoping to stimulate new research in individual animal identification from images, we propose to formulate the problem as the human-machine CONTINUAL CURATION *of images and animal identities. This is an open world recognition problem [4], where most new animals enter the system after its algorithms are initially trained and deployed.* CONTINUAL CURATION*, as defined here, requires (1) an improvement in the effectiveness of current recognition methods, (2) a pairwise verification algorithm that allows the possibility of no decision, and (3) an algorithmic decision mechanism that seeks human input to guide the curation process. Error metrics must evaluate the ability of recognition algorithms to identify not only animals that have been seen just once or twice but also recognize new animals not in the database. An important measure of overall system performance is accuracy as a function of the amount of human input required.*

1. Introduction

The challenging problem of automatically identifying unique animals from images (animal ID) is of growing importance in biology, ecology and conservation. See [25] for a recent review. Applications in animal ID range from population surveys to studying the behavior and movements of single animals. This problem has received more attention by computer vision community as well, as techniques have been developed for a wide variety of species, including small birds [11], penguins, primates [10, 12, 24], ungulates [9], large predators [8, 17], elephants [26], manta rays [18] sharks [1, 14] and whales [26].

Most work in the computer vision literature on animal ID treats the problem as an extension of the object and face recognition problem. In particular, the database of individuals is assumed to largely be closed, the problem is posed as one of retrieval, the primary goal is rank-ordering of potential IDs for each query, and few experiments are presented on handling novel identities. While this problem setting is important for some applications – and improvements in retrieval algorithm performance are still needed – we argue that a broader formulation, to which we refer as CONTINUAL CURATION, is needed for the animal ID problem. This is a problem where (1) most animals are added to the system after it is initially trained and deployed, (2) when these new animals are added it is not known that they are truly novel, and (3) individuals often must be re-identified very soon (if not immediately) after being added to the system.

Our formulation has a number of important implications for the development of animal ID algorithms that we discuss in this paper. We start by defining the problem in detail (Section 2). We then outline the algorithm pipeline we are developing for the REDACTED[1] system, and consider its relationship to associated problems and techniques in machine learning and computer vision (Section 3). Finally, we introduce several performance metrics (Section 4). While our system remains a work in progress, our goal is to stimulate new work in the field, leading to the development and sharing of datasets, algorithms, performance metrics and software throughout the community.

Before beginning, we must stress that our focus is on algorithmic issues. There remain equally important application issues surrounding the collection, protection and sharing of animal ID data and its derived metadata, especially for endangered or poached species. For brevity and focus, we leave those considerations for future discussions.

2. Animal ID and Continual Curation

The following observations about animal ID are based on real-world experience in using prototype systems for animal population monitoring and management.

1. New individuals are continuously added to the system, making animal ID an open world recognition problem [4]). Moreover, when new images are added it is not known which of them may show new individuals. Thus, any solution must recognize when an individual is new and must *quickly* learn to recognize it in subsequent images. These new individuals arrive in an unpredictable pattern.Thus, recognition of the new individual may be

[1] Name of author's public animal ID product redacted for peer review.

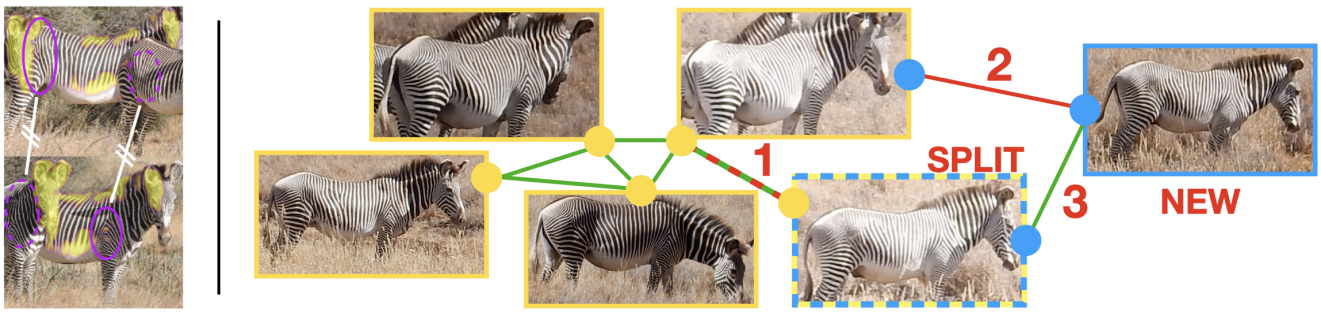

Figure 1. [LEFT] An example of two incomparable annotations of the same animal. [RIGHT] An example of a split decision. A database consists of 5 annotations for 1 ID (yellow) with positive matches. new image (blue) is added and gets both negative (2) and positive (3) decisions. To fix the problem, a split is done by making (1) negative and adding the blue/yellow annotation to the blue ID.

based on only one or a few examples. This is a fundamental challenge in recognition ([22]).

2. The distribution of animal sightings is typically heavily skewed, with many individuals appearing once in the database. For example, in a large whale shark dataset about 45% of over 10,000 individuals were seen once, while in a large Grevy's zebra census, 31% of over 2,000 individuals were only seen once.
3. Humans are an inherent part of the curation process, as training data annotators as well as the ultimate reviewers. Moreover, it is expected and often required that the curation decisions will include human judgement.
4. Humans typically find it challenging, given an image, to identify the animal in the image from a large animal population. (This is not necessarily true of small populations like the few rhinos on a conservancy.) Instead, humans are good at looking at two images and determining whether or not they show the same animal. The implications of this are quite important. Fundamentally, working without intelligent tools, a human would need to compare each query image to a representative image from each individual animal to decide whether it is new. The manual effort therefore grows quadratically with population size, leading to limits on the size of training sets or accuracy.
5. Metadata about IDs is rarely available, especially compared to face recognition datasets where identifying information can be gleaned from government records, surrounding text, and social media content [15].
6. Mistakes will enter the curation results. These are naturally caused by both imperfect algorithms and fallible human reviewers, but this does not tell the whole story. The vagaries of the data are often an important part of the issue. For example, minor differences in viewpoint and lighting may make two pictures that show the same individual look different. A third image may eventually be introduced that matches reliably against each. Animal ID algorithms must discover this and adjust, leading to the merge of previously distinct animals. Conversely, adding a new image may uncover the need to "split" of an animal in the database. See Figure 1 (right).
7. Not all animals depicted in an image are identifiable. This can be caused by the pose of the animal, occlusions from vegetation or other animals, poor image quality, or in sufficient resolution.See Figure 1 (left). *Identifiability* is not an absolute distinction, however, meaning that different humans and algorithms may produce different ID decisions for the same image.
8. There is an inherent trade-off across the effort of manual curation, the completeness of data, and the accuracy of curation [3]. If achieving near-perfect curation at all times is a central goal then the required human effort must increase to compensate for any potential failures of the underlying algorithms. Placing stricter controls — through preliminary detection algorithms or manual filtering — on which images become candidates for identification, could allow algorithms to perform better statistically and reduce the human effort. This is appropriate if rapid, low-cost population censusing is the goal. On the other hand, these controls may be loosened if extracting social networks or tracking few individuals is the goal.
9. In part due to the limitations of human annotators, when dealing with moderate sized populations, most individuals will enter the system as "new" rather than as part of an initial training set. Thus, there is no real sense of a training phase based on curated data followed by a testing phase where the system is used.

These factors lead us to an expanded view of the animal ID problem as CONTINUAL CURATION, where, starting from an often-small initial training set, each new set of images added to the database can show a new animal, can lead to the discovery and correction of mistakes, may require a series of human decisions, and will require the system to adjust itself for the next round of additions. More formally, we define the problem of CONTINUAL CURATION as, given a stream of images from an open set of animal identities, the goal is to continuously maintain the identification of ani-

mals in the images, allowing for human decisions.

3. Ideas and Approaches

We are designing the next generation of REDACTED to address individual animal identification as a CONTINUAL CURATION problem. This section describes the major components of the design and considers related work in computer vision and machine learning. This latter discussion is likely incomplete and we look forward to suggestions and discussion from the community.

3.1. Components of a Solution

Processing starts with a set of one or more query images and a database of images, annotations, and relationships. An *annotation* is the region surrounding a detected individual in an image, together with a species label and other attributes. *Relationships* between annotations indicate whether or not they show the same individual. The database, including the relationships between annotations, is considered dynamic and may change at any time.

The following are the processing steps for sets of query images:

1. **Detection**: Each image is processed to find annotations. Instance segmentation may be applied as well where sufficient training data are available. Details of this step, while important, are not considered here, but see [21].
2. **Filtering**: Annotations are filtered according to species, quality, and viewpoint. The goal is to ensure that distinguishing information appears in each annotation. They may be filtered further to ensure that they all show roughly the same coverage of the distinguishing features – e.g. both shoulder and hip on a zebra.
3. **Ranking**: each query annotation is matched against other query annotations and against the database to produce potential matching animals. (Each query annotation starts with its own unique, temporary name.) These are then rank-ordered, giving output typical of an object recognition or human face ID algorithm.
4. **Verification**: Two potentially matching annotations are evaluated to determine whether they show the same animal, show different animals, or there is insufficient information to tell – for example if they show non-overlapping views of distinguishing features. This "incomparable" label is especially important when there are few annotations per individual, avoiding premature mistakes, allowing tradeoff between accuracy and uncertainty. Adding the incomparable label means that verification decisions must be based on more than just examining distances in a latent embedding space since distances for incomparable pairs of annotations are unlikely to be meaningful, especially when there are few annotations for an individual.
5. **Decision**: the information provided by the detection, filtering, ranking, and verification algorithms must be combined into identity decisions. Given the imperfect nature of the algorithms, human input is needed as a guide. The human could be given complete control starting from the ranking, which is appropriate for smaller, closed populations. For larger populations and more frequent queries, some level of automation is necessary to coordinate information and discover inconsistencies. We treat the problem as one of dynamic, interactive clustering [2], where each cluster should contain all annotations from a single animal, but these clustering decisions can change dynamically as new input is provided. An important consideration is to determine what new information to seek. Based on the discussion above, for humans this must be in the form of decisions about whether or not two annotations (or a small set of annotations) show the same animal. In essence this functions as inserting a human as an alternative to algorithmic verification algorithms, particularly under uncertainty.

3.2. Relation to Common Problems in Machine Learning and Computer Vision

The following is a brief discussion of the overlap between the CONTINUAL CURATION problem and the relevant computer vision and machine learning literature. The detection problem dovetails nicely with work on the broader detection problem, with an emphasis on overlap, viewpoint, quality, and occlusion [21]. For filtering, characterizing identifiability is the most significant open question.

In ranking and verification, the most promising direction – and the one we are exploring – is semi-supervised and self-supervised learning. Recent work has shown marked improvements in standard classification problems through the use of contrastive loss together with heavily augmented positive samples and simultaneous optimization over large numbers of negative samples [5, 6, 7, 13]. This has been extended to a combination of supervised and unsupervised learning, which has shown experimental performance improvements over either alone [16]. On the other hand, continued work on backbone architectures and training loss functions does not seem poised for breakthroughs of significance to animal ID. For example, experimental evidence in [19] show minimal performance differences between different loss functions once the backbone architecture is fixed.

Other directions of investigation are important for the ranking step. On a surprising number of species, traditional methods based on keypoint matching have proven effective [9]. In many cases they may be used to bootstrap training data for deep learning recognition algorithms. Alternatively, keypoint-based methods may be applied in combination with deep learning methods where such methods are least-likely to succeed: novel individuals and few sightings [23]. Also interestingly is the idea of learning features from annotations without ID labels. This has shown success in

learning to detect dorsal fin edges on dolphins and other cetaceans, humpback flukes and elephant ears [26]. Descriptors for matching are then computed from outlines. Extracting training data requires significantly more work per image than bounding box labeling, but avoids the combinatorial problem of manual matching and, in principle, can even be applied where no IDs are available.

The verification problem posed here is a twist on traditional verification problems – most notably for face ID [20] – because of our introduction of the "incomparable" option as a three-way classification problem. Interestingly, high-confidence negative decisions may provide additional data input for contrastive loss functions, or help reduce the number of verification pairs a human must examine.

Finally, the decision problem is perhaps most closely related to interactive clustering [2]. The key addition to the problem is that algorithmic guidance is needed to make decisions about which annotations to show to humans, focusing the human effort. The alternative, typical of interactive clustering, is manual inspection of all results. This is labor intensive and does not scale with the size of the datasets.

4. Developing Performance Metrics

In suggesting ideas for experimental protocols for CONTINUAL CURATION, we focus on three areas: the performance of the ranking and verification algorithms, the overall system performance, and the amount of human effort. We must work from datasets that have already been curated, and re-create or simulate the conditions under which they were added to the system. (As our tools develop we will be able to curate increasingly larger sets.) Given M annotations, we assume m are already labeled and the remaining $M - m$ are added.

Measuring the performance of a verification algorithm is the most straightforward, requiring only that we have sufficient examples to handle the expected imbalance between positive, negative, and incomparable pairs. Judgment of incomparable is a human decision. Evaluating performance for ranking and for latent space embedding requires splitting the data so that among the $M-m$ test annotations some individuals are modeled as new, some individuals have been added since training completed, and the rest were used in training. Important for the latter two is the number of times an animal has previously been seen.

Turning to the judgment of overall performance for identifying M annotations, we work with clustering measures, where a cluster is a group of annotations determined to be the same animal. We assume a relatively large number of clusters, most or which are relatively small ($O(1)$ in size). There is wide variety of clustering performance measures, including precision and recall.These measures are dominated by large clusters. A simpler measure, well-suited to small clusters, is the number of extracted clusters that are also a ground-truth clusters, and the number of ground-truth clusters that are also found in the extracted clusters. These are measures of whether the clusters correctly represent the true animal ID. We propose the geometric mean of these two as a summary statistic. A final simple measure of accuracy is the actual number of clusters extracted. This is appropriate for population surveys where the primary goal is counting.

Human decisions are necessary to bridge the gap between the results of imperfect identification algorithms and accurate curation, while engendering trust in the engineered systems. Therefore, one important measure is the level of accuracy as a function of human effort. The primary human effort we propose to measure is the number of verification decisions made manually. Since it is impractical to involve human decision-making in repeated experiments involving thousands of images, we suggest a simulation: algorithms request human verification decisions from an oracle, which consults the ground truth data and returns the correct answer with high probability and an intentionally incorrect answer otherwise. By varying this probability we can analyze both the impact and the tradeoff of human effort and accuracy.

These measures are straightforward to implement, but an important challenge in using them is having sufficient ID data. There are some species such as zebras, humpback whales, whale sharks and others[2] for which this may be possible and we hope to release datasets in the future. We also encourage the community to consider ways to contribute datasets for the study of CONTINUAL CURATION.

5. Discussion

We have presented an expanded view of the animal ID problem that we refer to as CONTINUAL CURATION. It is an open world problem featuring frequent addition of unknown new individuals through query images. It also features skewed distributions of animal sightings with many animals seen only once or just a few times, yet must be recognized in future images. These present significant challenges to current recognition technologies. They also necessitate an algorithmic decision-making mechanism based on interactive clustering. Given the challenges of the problem, and the need for accurate results, human decisions must be an integral part of the solution. In summary, we propose to measure overall system effectiveness by the accuracy of curation as a function of human effort.

Important next steps include implementation of the proposed experimental protocols and performance metrics, organization and dissemination of datasets, and measuring the effectiveness of current techniques to establish a baseline. We hope rapid progress will soon follow and we look forward to extensive discussion and critical feedback.

[2]LILA BC - http://lila.science/datasets/

References

[1] Z. Arzoumanian, J. Holmberg, and B. Norman. An astronomical pattern-matching algorithm for computer-aided identification of whale sharks rhincodon typus. *Journal of Applied Ecology*, 42(6):999–1011, 2005. 1

[2] J. Bae, T. Helldin, M. Riveiro, S. Nowaczyk, M.-R. Bouguelia, and G. Falkman. Interactive clustering: a comprehensive review. *ACM Computing Surveys (CSUR)*, 53(1):1–39, 2020. 3, 4

[3] A. Beard. Computer-assisted human annotation for animal identification. M.S. thesis, Department of Computer Science, RPI, 2020. 2

[4] A. Bendale and T. Boult. Towards open world recognition. In *Proc. IEEE CVPR*, pages 1893–1902, 2015. 1

[5] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations, 2020. 3

[6] T. Chen, S. Kornblith, K. Swersky, M. Norouzi, and G. Hinton. Big self-supervised models are strong semi-supervised learners. In *NeurIPS*, 2020. 3

[7] X. Chen, H. Fan, R. Girshick, and K. He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 3

[8] M. Clapham, E. Miller, M. Nguyen, and C. T. Darimont. Automated facial recognition for wildlife that lack unique markings: A deep learning approach for brown bears. *Ecology and Evolution*, 10(23):12883–12892, 2020. 1

[9] J. P. Crall, C. V. Stewart, T. Y. Berger-Wolf, D. I. Rubenstein, and S. R. Sundaresan. Hotspotter—patterned species instance recognition. In *2013 IEEE workshop on applications of computer vision (WACV)*, pages 2¡30–237. IEEE, 2013. 1, 3

[10] D. Deb, S. Wiper, A. Russo, S. Gong, Y. Shi, C. Tymoszek, and A. Jain. Face recognition: Primates in the wild, 2018. 1

[11] A. C. Ferreira, L. R. Silva, F. Renna, H. B. Brandl, J. P. Renoult, D. R. Farine, R. Covas, and C. Doutrelant. Deep learning-based methods for individual recognition in small birds. *Methods in Ecology and Evolution*, 11(9):1072–1085, 2020. 1

[12] A. Freytag, E. Rodner, M. Simon, A. Loos, H. S. Kühl, and J. Denzler. Chimpanzee Faces in the Wild: Log-Euclidean CNNs for Predicting Identities and Attributes of Primates. In B. Rosenhahn and B. Andres, editors, *Pattern Recognition*, volume 9796, pages 51–63. Springer International Publishing, Cham, 2016. 1

[13] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 3

[14] B. Hughes and T. Burghardt. Automated visual fin identification of individual great white sharks. *International Journal of Computer Vision*, 122(3):542–557, 2017. 1

[15] I. Kemelmacher-Shlizerman, S. M. Seitz, D. Miller, and E. Brossard. The MegaFace Benchmark: 1 Million Faces for Recognition at Scale. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4873–4882, Las Vegas, NV, USA, June 2016. IEEE. 2

[16] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan. Supervised contrastive learning. *arXiv preprint arXiv:2004.11362*, 2020. 3

[17] S. Li, J. Li, W. Lin, and H. Tang. Amur Tiger Re-identification in the Wild. *arXiv:1906.05586 [cs]*, June 2019. 1

[18] O. Moskvyak, F. Maire, A. O. Armstrong, F. Dayoub, and M. Baktashmotlagh. Robust re-identification of manta rays from natural markings by learning pose invariant embeddings. *arXiv preprint arXiv:1902.10847*, 2019. 1

[19] K. Musgrave, S. Belongie, and S.-N. Lim. A metric learning reality check. In *European Conference on Computer Vision*, pages 681–699. Springer, 2020. 3

[20] M. H. Papesh. Photo id verification remains challenging despite years of practice. *Cognitive Research: Principles and Implications*, 3(1):1–9, 2018. 4

[21] J. Parham, C. Stewart, J. Crall, D. Rubenstein, J. Holmberg, and T. Berger-Wolf. An Animal Detection Pipeline for Identification. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1075–1083, Lake Tahoe, NV, Mar. 2018. IEEE. 3

[22] W. J. Scheirer, A. de Rezende Rocha, A. Sapkota, and T. E. Boult. Toward open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(7):1757–1772, 2012. 2

[23] S. Schneider, G. W. Taylor, S. S. Linquist, and S. C. Kremer. Similarity learning networks for animal individual re-identification - beyond the capabilities of a human observer. *CoRR*, abs/1902.09324, 2019. 3

[24] D. Schofield, A. Nagrani, A. Zisserman, M. Hayashi, T. Matsuzawa, D. Biro, and S. Carvalho. Chimpanzee face recognition from videos in the wild using deep learning. *Science advances*, 5(9):eaaw0736, 2019. 1

[25] M. Vidal, N. Wolf, B. Rosenberg, B. P. Harris, and A. Mathis. Perspectives on individual animal identification from biology and computer vision, 2021. 1

[26] H. Weideman, C. Stewart, J. Parham, J. Holmberg, K. Flynn, J. Calambokidis, D. B. Paul, A. Bedetti, M. Henley, F. Pope, and J. Lepirei. Extracting identifying contours for african elephants and humpback whales using a learned appearance model. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, March 2020. 1, 4