Interactive Image Retrieval for Biodiversity Research

Alexander Freytag[1,2], Alena Schadt[1], and Joachim Denzler[1,2]

[1]Computer Vision Group, Friedrich Schiller University Jena, Germany
[2]Michael Stifel Center Jena, Germany

Abstract. On a daily basis, experts in biodiversity research are confronted with the challenging task of classifying individuals to build statistics over their distributions, their habitats, or the overall biodiversity. While the number of species is vast, experts with affordable time-budgets are rare. Image retrieval approaches could greatly assist experts: when new images are captured, a list of visually similar and previously collected individuals could be returned for further comparison. Following this observation, we start by transferring latest image retrieval techniques to biodiversity scenarios. We then propose to additionally incorporate an expert's knowledge into this process by allowing him to select must-have-regions. The obtained annotations are used to train exemplar-models for region detection. Detection scores efficiently computed with convolutions are finally fused with an initial ranking to reflect both sources of information, global and local aspects. The resulting approach received highly positive feedback from several application experts. On datasets for butterfly and bird identification, we quantitatively proof the benefit of including expert-feedback resulting in gains of accuracy up to 25% and we extensively discuss current limitations and further research directions.

1 Introduction

In biodiversity research, experts are confronted with a growing amount of collected images which build the foundation for statistics over distributions of species, their habitats, or the overall biodiversity in ecosystems. Within this challenging process, classification of individuals is commonly done using field guides and by comparing the current object of investigation against known classes, thereby checking for the presence of unique characteristics (*e.g.,* a red dotted neck or a characteristically colored wing). Common image retrieval techniques, *e.g.,* [26, 8, 34, 27, 2, 4, 37, 5], could greatly assist in this process by suggesting visually similar genera for further inspection to an expert. Simply applying these techniques to biodiversity scenarios, however, does not necessarily lead to satisfying results. One reason is that species, while visually similar on a global scale, often show significant differences in small and localized details which are easily missed. Furthermore, locations of discriminative details significantly differ between categories, which requires experts to investigate different sets of parts depending on the currently faced individual. In this paper, we therefore present an approach to improve existing image retrieval techniques by incorporating expert feedback about must-have-regions into the retrieval process. A visualization of the underlying idea is given in Fig. 1.

We will build on neural codes [4] as baseline, a recently introduced technique for image retrieval using activations of convolutional neural network architectures. Thereby,

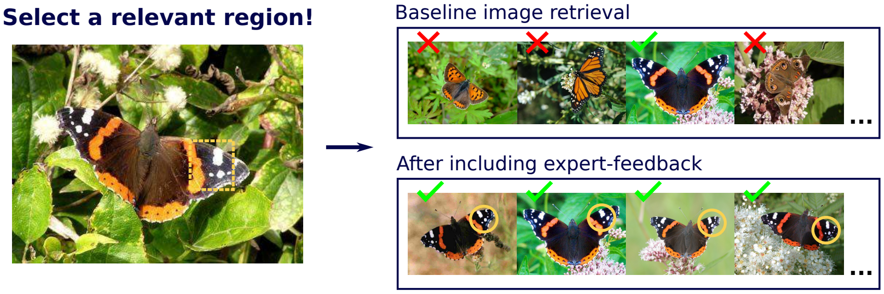


Fig. 1: Image retrieval techniques can assist biodiversity researchers by filtering collected datasets for individuals visually similar to an unseen object (*left and top row*). We present how an expert-in-the-loop can improve this baseline by selecting a must-have-region (*e.g., the dashed rectangle*). By training region-specific detection models, we can detect these regions in training images (*indicated as circles in bottom row*) to verify and update the initial ranking.

handing over a query image results in similarity scores for all previously collected training examples. Based on an image region specified by an expert, we then learn a detection model from only this single positive example following recent results for exemplar-models in patch discovery [30, 19, 12]. Efficient evaluations of the detection model on all training images result in a second score indicating the presence of the selected region. We finally update the ranking by fusing both results. In consequence, we obtain a list of visually similar individuals which additionally exhibit parts similar to the selected region. For the sake of quantitative results, we require ground truth labels of training images in the presented evaluations. However, our approach does *not* rely on class annotations at all and can thus be applied even in unsupervised tasks where no (machine-accessible) class information is available. Furthermore, the resulting approach runs within seconds on standard hardware and is thus also applicable for large image collections.

In the remainder of this paper, we first give a short review on state-of-the-art in image retrieval (Section 2) and then introduce our approach for interactive image retrieval in Section 3. Quantitative analyses of the resulting system are presented in Section 4 on computer vision datasets related to biodiversity applications. Depending on the tackled scenario, we report significant accuracy gains but also show and discuss limitations of the current approach and resulting future research directions. A short summary concludes the paper.

2 Image Retrieval in a Nutshell

Retrieving visually similar images given a newly captured query, *i.e.,* the task we motivated in the last section, is the central topic of an entire research area which is commonly referred to as *content-based image retrieval*. In more then 20 years of research, a variety of approaches has been developed, *e.g.,* [15, 31, 26, 8, 27, 2, 1, 4, 37, 7, 5]. While differing in algorithmic details and required assumptions, the underlying pipeline mainly consists of three steps: (i) representing known images and organizing representations in a search structure, (ii) computing representations for a new query image, and (iii) matching of representations to build a ranked list from which top-ranked images are returned. A great amount of research and engineering art went into carefully designing and implementing all three steps. However, we noticed two crucial issues for image retrieval in biodiversity applications. First of all, latest findings from the image retrieval community are yet to be transferred to remaining areas of application. While this is partly successful, *e.g.,* in medical scenarios [38], we found in several discussions that this process works rather poorly in biodiversity research. Besides, we found that off-the-shelf retrieval algorithms are often not perfectly feasible for biodiversity applications. This observation mainly arises from the fact that a large fraction of retrieval algorithms aim at finding images of exactly the same object as the query [31, 26, 8, 27, 2, 4, 37, 5]. In this paper, we are instead interested in retrieving known individuals similar but not identical to a previously *unseen* sample. This task is by far not novel, and an entire sub-field known as *category retrieval* tackles this challenging problem by modeling or learning the occurring intra-category variances (*e.g.,* see [7, 4, 5] for latest impressive results as well as [22] for an application to plant species identification scenarios). While we received already promising feedback by simply applying image retrieval techniques in biodiversity scenarios, the resulting framework was often found to be too static. Instead, the possibility for selecting must-have-regions, *e.g.,* a unique wing pattern, was often desired. In terms of computer vision, we thus seek for distance metrics which are user- and exemplar-specific and interpretable.

A research area similar in spirit is local learning, where known images most similar to a test image are retrieved to then learn representations and models from those similar images only [40, 13, 32, 17, 12, 7, 39]. In the focus of this paper, we instead leave the decision process to the expert, but aim at providing him with a set of relevant images as helpful as possible and further allow him to interactively refine the query results.

The only related work we a aware of is [6] which allows to select outputs of an unsupervised segmentation for query refinement. We instead propose a more intuitive and precise technique for providing feedback as shown next.

3 Interactive Selection of Regions of Focus

Let us now introduce the aforementioned technique for interactive image retrieval. As a result of several discussions, we found that the selection of rectangles as must-have-regions for the current query is an intuitive, simple, and yet powerful way for receiving feedback from an application expert. To integrate this information into the retrieval process, we provide solutions for three questions:

i. how to train a detection model f from a single positive example,
ii. how to efficiently evaluate the model on all training images, and
iii. how to integrate detection responses into the process of image ranking.

An overview of the resulting approach is shown in Fig. 2.

Exemplar-models for Region Detection. Determining the existence or absence of a selected region in a collected training set can be done most easily by casting it as a part

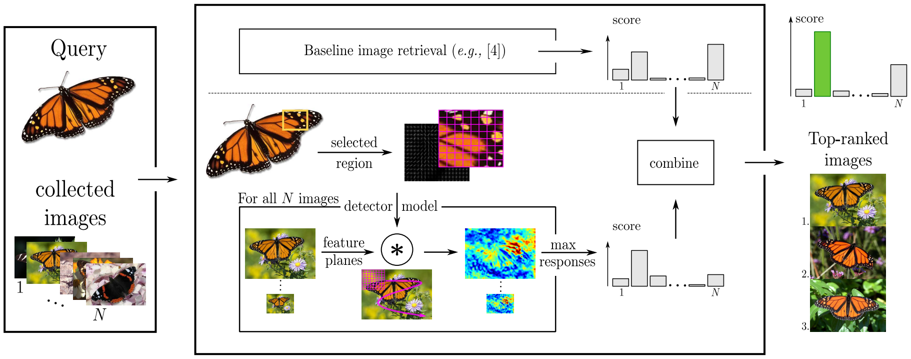


Fig. 2: Overview of our proposed approach for interactive image retrieval. Figure is best viewed in color and by zooming in. See text for details.

detection task. Thus, we aim at training a detection model from a single positive example and virtually everything else as negative data. While this task appears cumbersome on first glance, exemplar-models such as Exemplar-SVM [24] or Exemplar-LDA [16, 19] provide an elegant solution and have been found useful for learning patch detectors from a single positive example [30, 19, 12]. Unfortunately, training of Exemplar-SVMs involves computationally expensive hard negative mining, In contrast, Exemplar-LDA models can be trained highly efficiently since the majority of computations is done only once in an offline stage. Since we are interested in fast responses after an expert selected a region, we thus follow [19, 12] and apply LDA models as region detectors. In consequence, distributions of (the single) positive and all negative examples are assumed to be Gaussian with a shared covariance matrix $\boldsymbol{\Sigma}_0$ and mean vectors $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_{-1}$, respectively. Although this assumption might be far from being perfectly correct, it offers two advantages: (i) a discriminative linear separation of positive and negative data

$$\boldsymbol{w}_{\mathrm{LDA}} = \boldsymbol{\Sigma}_0^{-1} \left(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_{-1} \right) \tag{1}$$

with (ii) fast training and model evaluations [16]. Furthermore, we can additionally view Eq. (1) as de-correlated nearest class mean [25] and it is known that reducing correlations in feature cells is beneficial for detection tasks [16, 12]. We thus only need to compute $\boldsymbol{\Sigma}_0$ and $\boldsymbol{\mu}_{-1}$ from all possible locations, aspect ratios, and scales in all training samples once [16]. In practice, we add a scaled identity matrix $\boldsymbol{\Sigma}_0 + \sigma_n^2 \boldsymbol{I}$ to increase numerical stability. During the interaction process, it only remains to solve the linear equation system in Eq. (1) to obtain the desired detection model.

Efficient Convolutions for Region Detection. To reliably detect the selected region in training images, we need to densely evaluate the learned detector, *i.e.,* on all possible locations and scales. Let therefore $\boldsymbol{x}$ denote the feature vector extracted from a single position and scale. For a linear detector as in Eq. (1), the response on $\boldsymbol{x}$ is computed as additive combination of dimension-specific similarity scores (ignoring offsets for simplicity of notation):

$$f(\boldsymbol{x}) = \langle \boldsymbol{w}_{\mathrm{LDA}}, \boldsymbol{x} \rangle = \sum_{d=1}^{D} \boldsymbol{w}_{\mathrm{LDA}}(d) \cdot \boldsymbol{x}(d) \quad . \tag{2}$$

Evaluation of f on all possible locations can then be done in a sliding window manner by computing Eq. (2) for densely extracted features. In this case, we can also change the order of computations and can equivalently compute Eq. (2) by adding D convolutions of 1×1 filters with corresponding feature planes. As required later, this also holds if $\boldsymbol{x}$ follows a spatial tiling composed of $T \times T$ cells with D_C feature dimensions per cell (thus, $D = T \cdot T \cdot D_C$ in Eq. (2)). For prominent examples, *e.g.,* Spatial Pyramid Match Kernels [23] or HOG [11] for detection tasks, Eq. (2) translates to

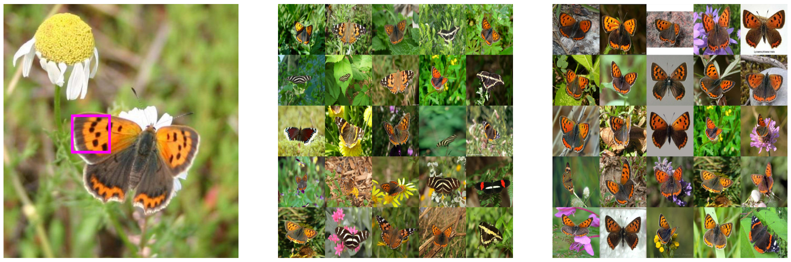
$$f(\boldsymbol{x}) = \sum_{d_i, d_j = 1}^{T} \sum_{c=1}^{D_C} \boldsymbol{w}_{\mathrm{LDA}}\left((d_j T + d_i) D_C + c\right) \cdot \boldsymbol{x}\left((d_j T + d_i) D_C + c\right) \quad . \tag{3}$$

Again, we can exchange order of summations which leads to adding results of D_C convolutions of $T \times T$ filters with corresponding feature planes. By computing feature planes for all training images in an offline step, we can efficiently detect selected regions and reduce an expert's idle times to a minimum.

Fusion of Complementary Retrieval Scores. Given the previous steps, it now remains to combine detection results with the previously obtained ranking of the baseline retrieval system. As commonly done in object detection, we perform max-pooling over response maps from all scales and return the largest detection score for each image. Scores are linearly normalized into $[0, 1]$ to maintain their relative ordering and still allow for a well-defined range of outputs. Given results of baseline image retrieval and interactive selection, we now seek for examples with high scores reflected by both indicators. We therefore assume both sources of information to be complementary which justifies a simple product as combination rule [3, 20]. Note that the assumption of complementary information is indeed justifiable, since a baseline retrieval is concerned with coarse distinctions regarding the entire image. Instead, interactive selection explicitly neglects the majority of this information and searches for the remaining parts with arbitrarily different techniques. While we also experimented with other fusion techniques [20], we empirically found this strategy to be as simple as powerful. Combined scores are finally ordered and top-ranked results are returned. Putting all parts together, we obtain the framework as visualized in Fig. 2. The entire pipeline runs within seconds and allows experts to easily investigate results with different regions selected. A qualitative example is given in Fig. 3.

4 Experiments

By applying the previously introduced techniques to biodiversity tasks, we already obtained highly positive feedback from several experts which we took as a qualitative confirmation of our approach's usability. In this section, we additionally present quantitative results on two established computer vision datasets to analyze benefits, limitations, and future research directions.

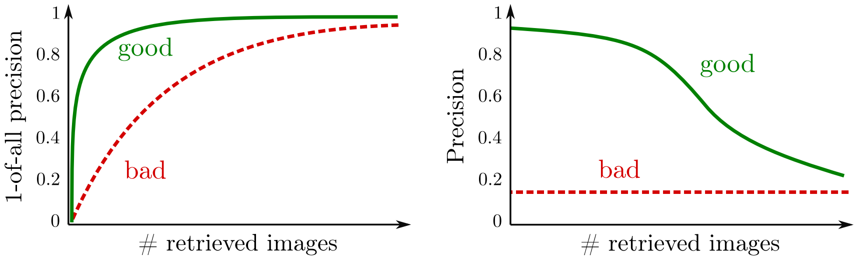


Query with selection Baseline only Combined

Fig. 3: Visualizing the influence of expert feedback. Figure is best viewed in color.

4.1 Evaluation Criteria

To evaluate success of an image retrieval technique, a variety of different criteria have been put forward and the presumably most common measure is mean average precision (mAP) [26, 8, 27, 2, 4, 37, 5] based on precision and recall. When returning k top-ranked images, *precision* refers to the relative number of correctly retrieved images, *i.e.,* $\frac{\#\text{true positives}}{k}$, whereas *recall* denotes the number of correctly retrieved images relative to the absolute number of known positive examples, *i.e.,* $\frac{\#\text{true positives}}{\#\text{known positives}}$. Computing mAP is then done by plotting recall against accuracy individually for each possible category and averaging areas under the resulting curves. While mAP is excellent for evaluating an image retrieval system's performance over the entire range of possible working points, *i.e.,* different trade-offs between precision and recall, we observed that the majority of possible working points is not feasible in practice. Instead, application constraints often render high recall values as an irrelevant measure of quality. According to our experience, an application expert is in fact not interested in a supporting tool with perfect recall which returns almost all known images – only to not miss a single correct one. Instead, he is usually interested in inspecting just a small set of retrieved images, and this retrieved set should exhibit several properties. Interestingly, we also observed that these properties vary over task and expert, *e.g.,* experienced researchers are usually interested in inspecting visually similar examples to then make a final decision on their own. In this case, retrieving at least one example of the correct category is often sufficient which we refer to as 1*-of-all precision*. Less experienced researchers, though, often base their decision on relative frequencies of returned categories. In these cases, as many retrieved images as possible should be of the correct category, *i.e.,* a high precision matters. While several papers followed the second evaluation, *e.g.,* [37, 7], we are not aware of any work applying the first principle, which however was found to be a useful criterion for application experts. We will see later that both criteria can cover orthogonal aspects of a system's performance and thus should be considered side by side. Both criteria are illustrated in Fig. 4 visualizing results for a strong and a poor retrieval system.



(a) 1-of-all precision

(b) Precision

Fig. 4: Illustrating different criteria for evaluating image retrieval accuracy.

4.2 Datasets for Illustrating Biodiversity Applications

So far, we are not aware of any biodiversity dataset publicly available for computer vision researchers. To still allow for quantitative evaluations, we present experimental results on two datasets established in the computer vision community which cover areas of investigation relevant for biodiversity researchers. In the following, we give a short overview on both datasets.

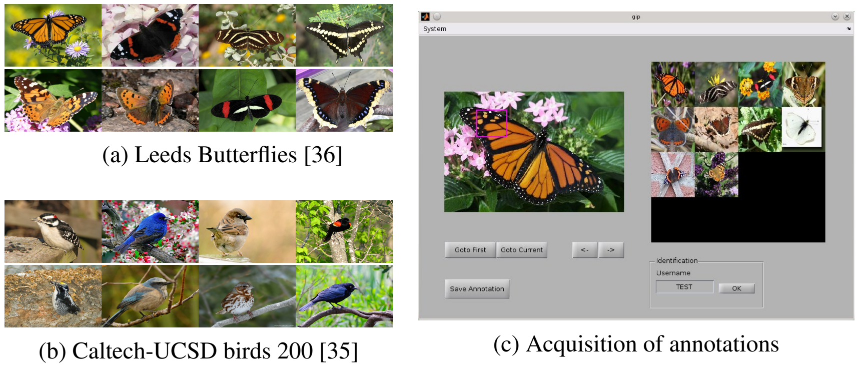
Leeds – Identifying Butterflies. The Leeds Butterfly dataset [36] contains 832 images of butterflies captured in a natural environment. It covers ten distinct butterfly species with 55 to 100 images per category. Exemplary individuals of eight species are shown in Fig. 5a.

CUB2011 – Recognizing Birds. The Caltech-UCSD Birds-200-2011 (CUB200) [35] dataset covers 200 bird species native in North America. The provided dataset contains $11{,}788$ individuals which are split in train and test set of approximately same size. We also conducted experiments on the frequently used subset (CUB14) by [10] which contains 14 categories of warblers and woodpeckers with 817 images. Examples are shown in Fig. 5b. Noteworthy, category labels do not distinguish between male and female, nor between young and adult.

4.3 Experiments in a Butterfly Identification Scenario

We have already seen a qualitative example in Fig. 3. For a quantitative evaluation, we start with the previously introduced butterfly dataset Leeds [36].

Experimental Setup. As baseline retrieval technique, we apply neural codes by [4]. In detail, we use the AlexNet model [21] initially trained on ImageNet for general purpose feature extraction [18, 14]. Features of several layers are extracted using the Caffe toolbox [18] and we empirically found `conv3` to be well suited for our application. Since the dataset does not provide any part information, we asked six users to manually select a single region for each image which they rate as informative. Notably, the users' initial domain knowledge ranged from no knowledge at all to individual training for several weeks. For their guidance, we displayed individuals of each category as visualized in Fig. 5c. Following recent trends in fine-grained recognition [12], we represent



(a) Leeds Butterflies [36]

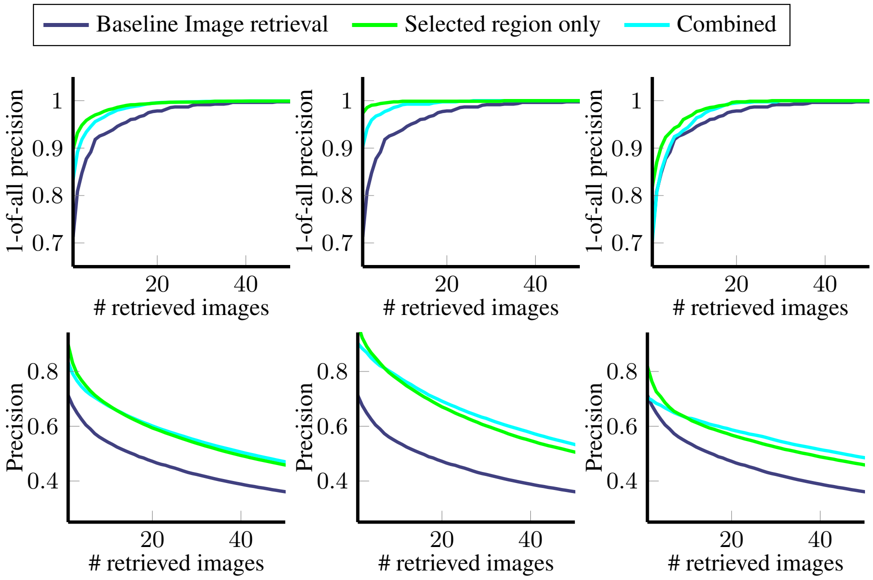
(b) Caltech-UCSD birds 200 [35]

(c) Acquisition of annotations

Fig. 5: *Left*: Examples of different species from datasets used in our evaluations showing butterflies [36] (*top*) and birds [35] (*bottom*). *Right*: GUI for acquisition of must-have-regions to allow for quantitative evaluations.

selected regions using histograms of oriented gradients (HOG) [11] and histograms of ColorNames (CN) [33] to capture color, texture, and shape. Spatial information is kept by tiling the selected region using a regular grid and extracting features in each cell separately [11]. HOG and CN features are extracted using publicly available source-code of [16] and [33]. We train exemplar-specific LDA models using the code provided by [12]. For evaluation, we follow the leave-one-out principle and exclude each image once from the training set to serve as query image. Precision and 1-of-all precision curves are finally averaged over all images and shown in Fig. 6.

Evaluation. When averaging over all users (Fig. 6a), we notice a significant increase in both precision and 1-of-all precision. Noteworthy, the accuracy with respect to the first retrieved image is increased from 71% to 90%. From our experience, this result is indeed remarkable given the already sophisticated performance obtained with neural codes as baseline technique. Furthermore, we note that experienced annotators can easily lift the retrieval accuracy to ranges significantly over 95% (Fig. 6b). However, even novices with little experience can add valuable information (Fig. 6c). We also observe that solely relying on outputs of detection models further boosts performance if k is extremely small but is inferior to combined results for larger retrieved sets. This behavior is plausible since images with extremely high detection scores are likely to contain the exact same pattern as the query. In contrast, medium scores likely result from examples of mixed-up categories which have a similar local pattern but are different at a global scale. Consequently, incorporating the baseline information can correct these cases. Interestingly, we also notice different trends when comparing precision and 1-of-all precision as a measure of accuracy. We therefore conclude that a decision for one evaluation strategy over the other should be based on the desired properties of the retrieved set of images.



(a) Average over users (b) Best expert (c) Worst novice

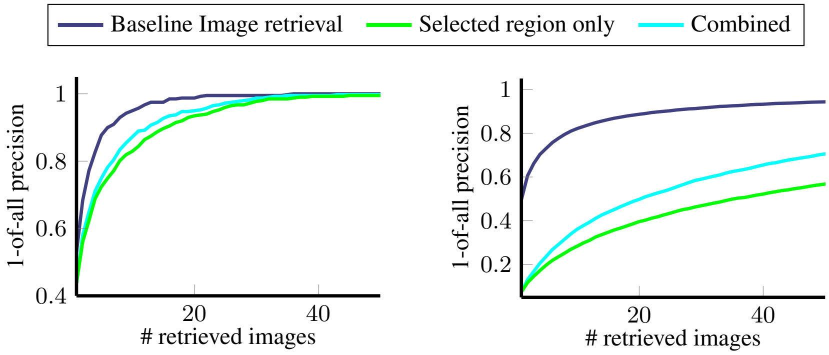
Fig. 6: Evaluating image retrieval with interactive feedback on the butterfly dataset Leeds [36]. Results are obtained from six annotators ranging from novices to experts.

4.4 Experiments in a Bird Recognition Task

In a second experiment, we evaluate limitations of our approach and further research directions using the previously introduced CUB200 bird dataset.

Experimental Setup. Following previous research [42, 12, 9], we crop images to the provided bounding box and apply the provided split in train and test images. We simulate region selection using provided part annotations for anchoring a squared region of width and height proportional to $\frac{1}{10}$ of the box's main diagonal. For verification, we additionally asked our most experienced annotator to mark head regions on the small subset. The remaining setup is identical to the previous experiment except that neural codes are extracted from `conv5`. Due to the lack of space, we only present results in Fig. 7 obtained from head regions which are known to be most discriminative [42].

Evaluation. In contrast to the superior results on Leeds, we notice that selecting a single region is too restrictive for bird recognition tasks and can even hurt retrieval accuracy. Notably, even our most experienced expert was not able to improve the accuracy (Fig. 7a). On CUB200, the accuracy induced by detection scores finally drops significantly (Fig. 7b) and thus goes along with the combined results. We attribute this observation to three reasons. First of all, captured bird images are highly diverse, both with respect to pose (parts are often occluded, thus, no model can be trained) and appearance (male vs. female, young vs. adult, label errors). Besides, the number of species is significantly larger which renders the task significantly more difficult. Finally, single parts



(a) CUB14

(b) CUB200

Fig. 7: Evaluating limitations and further research directions for the introduced interactive retrieval approach on bird recognition datasets by [35].

are often not visually discriminative with respect to different categories although their combination is. Since the usability of the current approach was confirmed in personal discussions, we can conclude several important directions for improvement. First of all, an extension to multiple selectable regions would be highly beneficial to specify parts which are only discriminative when appearing jointly. Besides, estimating the number of required annotations would be helpful for unexperienced researchers. Finally, providing relative positions of multiple parts and expressing their semantics would allow for more informative expert feedback.

5 Conclusions

In this paper, we introduced image retrieval techniques to assist in biodiversity research. Using neural codes as baseline, we then presented how to additionally incorporate expert feedback by interactively selecting must-have-regions. The provided information served for training of region-specific detection models which are efficiently evaluated on all training images with convolutions. Combining detection scores with baseline results finally allowed for verifying and updating the initial ranking. In a butterfly identification task, this intuitive way of providing feedback resulted in improved results for non-experts while more experienced users could even further boost the performance. The resulting approach is easy to use and already received highly positive feedback from several experts. In a last experiment, we evaluated limitations of our approach and discussed open research directions. As future work, we plan to incorporate relevance feedback [29, 41] which was suggested by medical experts. In addition, transferring our approach to different application areas, *e.g.,* retrieval of similar plants [22] or moths [28], could assists experts in other domains. Furthermore, replacing current region descriptions by efficiently computable CNN activations is likely beneficial. While their applicability is currently limited by hardware requirements, further progress in this field will allow for training even better region detection models.

References

1. Arandjelovic, R.: Advancing Large Scale Object Retrieval. Ph.D. thesis, University of Oxford (2013)
2. Arandjelovic, R., Zisserman, A.: Three things everyone should know to improve object retrieval. In: CVPR (2012)
3. Atrey, P.K., Hossain, M.A., El Saddik, A., Kankanhalli, M.S.: Multimodal fusion for multimedia analysis: a survey. Multimedia systems 16(6), 345–379 (2010)
4. Babenko, A., Slesarev, A., Chigorin, A., Lempitsky, V.: Neural codes for image retrieval. In: CVPR (2014)
5. Cao, X., Zhang, H., Guo, X., Liu, S., Chen, X.: Image retrieval and ranking via consistently reconstructing multi-attribute queries. In: ECCV (2014)
6. Carson, C., Belongie, S., Greenspan, H., Malik, J.: Blobworld: Image segmentation using expectation-maximization and its application to image querying. TPAMI 24(8), 1026–1038 (2002)
7. Chatfield, K., Simonyan, K., Zisserman, A.: Efficient on-the-fly category retrieval using convnets and gpus. In: ACCV (2014)
8. Chum, O., Philbin, J., Sivic, J., Isard, M., Zisserman, A.: Total recall: Automatic query expansion with a generative feature model for object retrieval. In: ICCV (2007)
9. Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., Darrell, T.: Decaf: A deep convolutional activation feature for generic visual recognition. In: ICML (2014)
10. Farrell, R., Oza, O., Zhang, N., Morariu, V.I., Darrell, T., Davis, L.S.: Birdlets: Subordinate categorization using volumetric primitives and pose-normalized appearance. In: ICCV (2011)
11. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part based models. TPAMI 32(9), 1627–1645 (2010)
12. Freytag, A., Rodner, E., Darrell, T., Denzler, J.: Exemplar-specific patch features for fine-grained recognition. In: GCPR. pp. 144–156 (2014)
13. Frome, A., Singer, Y., Sha, F., Malik, J.: Learning globally-consistent local distance functions for shape-based image retrieval and classification. In: ICCV. pp. 1–8 (2007)
14. Guadarrama, S., Rodner, E., Saenko, K., Zhang, N., Farrell, R., Donahue, J., Darrell, T.: Open-vocabulary object retrieval. Robotics: Science and Systems (2014)
15. Gudivada, V.N., Raghavan, V.V.: Content based image retrieval systems. IEEE Computer (Special Issue on Content-Based Image Retrieval Systems) 28(9), 18–22 (1995)
16. Hariharan, B., Malik, J., Ramanan, D.: Discriminative decorrelation for clustering and classfication. In: ECCV (2012)
17. Ionescu, R., Popescu, M., Grozea, C.: Local learning to improve bag of visual words model for facial expression recognition. In: ICML Workshop on Representation Learning (2013)
18. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: Convolutional architecture for fast feature embedding. arXiv preprint arXiv:1408.5093 (2014)
19. Juneja, M., Vedaldi, A., Jawahar, C., Zisserman, A.: Blocks that shout: Distinctive parts for scene classification. In: CVPR (2013)
20. Kittler, J., Hatef, M., Duin, R.P., Matas, J.: On combining classifiers. TPAMI 20(3), 226–239 (1998)
21. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: NIPS (2012)
22. Kumar, N., Belhumeur, P.N., Biswas, A., Jacobs, D.W., Kress, W.J., Lopez, I.C., Soares, J.V.: Leafsnap: A computer vision system for automatic plant species identification. In: ECCV. pp. 502–516 (2012)

23. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: CVPR. pp. 2169–2178 (2006)
24. Malisiewicz, T., Gupta, A., Efros, A.: Ensemble of exemplar-svms for object detection and beyond. In: ICCV (2011)
25. Mensink, T., Verbeek, J., Perronnin, F., Csurka, G.: Distance-based image classification: Generalizing to new classes at near-zero cost. TPAMI 35(11), 2624–2637 (2013)
26. Philbin, J., Chum, O., Isard, M., Sivic, J., Zisserman, A.: Object retrieval with large vocabularies and fast spatial matching. In: CVPR. pp. 1–8 (2007)
27. Philbin, J., Chum, O., Isard, M., Sivic, J., Zisserman, A.: Lost in quantization: Improving particular object retrieval in large scale image databases. In: CVPR. pp. 1–8 (2008)
28. Rodner, E., Simon, M., Brehm, G., Pietsch, S., Wgele, J.W., Denzler, J.: Fine-grained recognition datasets for biodiversity analysis. In: CVPR-WS (2015)
29. Rui, Y., Huang, T.S., Ortega, M., Mehrotra, S.: Relevance feedback: a power tool for interactive content-based image retrieval. IEEE Transactions on Circuits and Systems for Video Technology 8(5), 644–655 (1998)
30. Singh, S., Gupta, A., Efros, A.: Unsupervised discovery of mid-level discriminative patches. In: ECCV (2012)
31. Sivic, J., Zisserman, A.: Video google: A text retrieval approach to object matching in videos. In: ICCV. pp. 1470–1477 (2003)
32. Urtasun, R., Darrell, T.: Sparse probabilistic regression for activity-independent human pose inference. In: CVPR (2008)
33. Van De Weijer, J., Schmid, C., Verbeek, J., Larlus, D.: Learning color names for real-world applications. IEEE Transactions on Image Processing 18(7), 1512–1523 (2009)
34. Vogel, J., Schiele, B.: Semantic modeling of natural scenes for content-based image retrieval. IJCV 72(2), 133–157 (2007)
35. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The caltech-ucsd birds-200-2011 dataset. Tech. Rep. CNS-TR-2011-001, California Institute of Technology (2011)
36. Wang, J., Markert, K., Everingham, M.: Learning models for object recognition from natural language descriptions. In: BMVC (2009)
37. Wang, Q., Si, L., Zhang, D.: Learning to hash with partial tags: Exploring correlation between tags and hashing bits for large scale image retrieval. In: ECCV (2014)
38. Xu, X., Li, B.: Automatic classification and detection of clinically relevant images for diabetic retinopathy. Medical Imaging (2008)
39. Yu, A., Grauman, K.: Fine-grained visual comparisons with local learning. In: CVPR (2014)
40. Zhang, H., Berg, A.C., Maire, M., Malik, J.: Svm-knn: Discriminative nearest neighbor classification for visual category recognition. In: CVPR. pp. 2126–2136 (2006)
41. Zhang, L., Lin, F., Zhang, B.: Support vector machine learning for image retrieval. In: ICIP. pp. 721–724 (2001)
42. Zhang, N., Farrell, R., Iandola, F., Darrell, T.: Deformable part descriptors for fine-grained recognition and attribute prediction. In: ICCV (2013)