# Birds of a Feather Flock Together – Local Learning of Mid-level Representations for Fine-grained Recognition

Alexander Freytag*, Erik Rodner*, and Joachim Denzler

Computer Vision Group, Friedrich Schiller University Jena, Germany
{firstname.lastname}@uni-jena.de
http://www.inf-cv.uni-jena.de

**Abstract.** When facing a new object, the human ability for implicitly comparing it to a small set of known objects and analyzing similarities and differences is astonishing due to speed and performance. In this paper, we propose to follow this observation by *locally* learning mid-level representations and models for every unseen example. Thus, our approach is complementary to the standard procedure of training a single model in advance based on a universal representation, such as a single set of attributes. As an illustrating example, we evaluate our approach for fine-grained recognition of bird species. Experimental results show that already with a simple query function based on color and edge histograms, the resulting image-specific representations are at the same time compact (with respect to dimensionality) and informative (with respect to classification accuracy).

## 1 Local Learning of Mid-level Representations

**Local Learning** Do we need to know how to differentiate between 20 eagle species, when we currently spot a singing bird? Structuring data into easier sub-problems is known for decades, and usually referred to as divide-and-conquer. However, the majority of our recognition systems are build on flat classification models treating all classes alike. Unfortunately, huge models are strongly correlated with high model capacities, which again results in ill-posed learning problems for moderately sized training data. Although hierarchical models introduce flexibility here, underlying hierarchies are usually based on linguistic relations, *e.g.*, WordNet, which are not necessarily related to visual consistency. Thus, a hierarchy based on visual features seems a plausible solution here. Nonetheless, even such a scheme might still be too rigid for real-world scenarios. Intuitively, an optimal dividing strategy would adapt to every test sample and splits data according to query-specific criteria, *i.e.*, for every test sample, only 'related' training examples are considered for model training, which is known as *local learning*. An obvious benefit is that only features relevant for distinguishing the visually similar images are used for a classification, guiding model training towards the important aspects of related species. We transfer this idea to the challenging task of fine-grained recognition by learning representations and models from related images only. Our approach is visualized in Fig. 1.

* A. Freytag and E. Rodner were supported by a FIT scholarship from the DAAD.

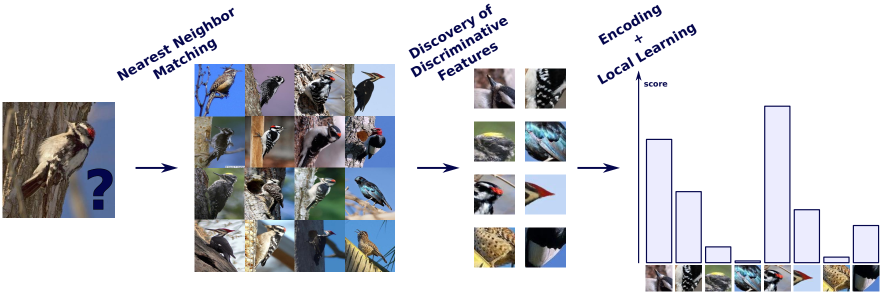

Nearest Neighbor Matching
Discovery of Discriminative Features
Encoding + Local Learning
score
?

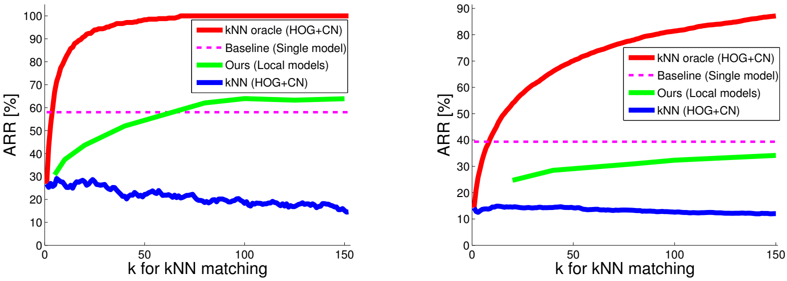
Fig. 1: Whereas common approaches train a single model with a universal, fixed representation, we propose to query for a given test image (*left*) its most similar training images (*middle left*). Learning of mid-level representations is done on the retrieved set only and thus results in parts relevant to differentiate highly similar objects (*middle right*). Query set and test example are then encoded with the discovered patches (*right*) and a local model is trained, which finally serves for classifying the query image.

In the early nineties, local learning was introduced in [2] and motivated by a suitable trade-off between model capacity and number of training examples, especially if training samples are non-uniformly distributed in space. Some rare exceptions followed this idea through the years, *e.g.*, [12] for image categorization, or [6] for recognizing facial expressions. One noteworthy byproduct of local learning techniques is the fact that models can be easily extended to previously unseen classes over time, making them especially favorable for lifelong learning scenarios [8].

**Learning of Mid-level Representations** Learning mid-level representations gained attention from our community within the last years. Usually, those representations are referred to as parts [1], attributes [3], or patches [9], and they are supposed to represent discriminative, reoccurring visual patterns in images of current interest. Although there is no necessity for involved semantics, many approaches explicitly mine for human interpretable attributes in various ways. Here, we skip a comprehensive review of current techniques, and exemplarily mention [9] which brought patch discovery back into focus of our field. Moreover, the local learning scheme we want to advertise here can be applied to any discovery scheme for mid-level features in principle. In contrast to previous approaches, our technique learns image representations *and* classification models for every test sample on the fly, which allows focusing on patches important to differentiate quite similar birds already observed (see Fig. 1).

## 2 Experiments on CUB-2011

**Experimental Setup** The CUB-2011 dataset [11] is a great playground to evaluate our approach, and we present experimental results using the whole dataset with all 200 classes as well as results on the frequently used 14 class subset. In order to discover a compact, informative, and image specific feature representation, we first query for every



(a) CUB-2011-14

(b) CUB-2011-200

Fig. 2: Accuracy of simple k-NN matching on CUB2011 for a perfect oracle, majority voting, and our approach for different sizes of the query set.

test sample its $K$ most similar training samples using global image representations based on combined HOG and color name (CN) features [10]. The retrieved set then serves for patch discovery, where any of the established techniques can be applied. In order to keep things simple, we applied a method developed in our group [4] with source code readily available[1].

**Evaluating the kNN Matching** When querying $K$ training images and training a model only from these samples, the resulting accuracy is obviously limited by the quality of the global matching: if the correct class is not present among the retrieved set, the learned model will always be wrong. Thus, we first evaluate the quality of matching based on globally extracted HOG and CN features. Results are given in Fig. 2a and Fig. 2b (red and blue line) for various sizes of $K$ on the 14 class subset and the full 200 class dataset, respectively. Here, an oracle reflects the upper bound of accuracy by counting every set with at least a single sample of the correct class as success. Interestingly, already a rather small set of $\sim$ 20 neighbors is sufficient for CUB-2011-14 to be far over state-of-the-art results, whereas CUB-2011-200 requires larger neighborhoods to guarantee the existence of the correct class among the retrieved set. Note that although the majority vote classification is unlikely to result in proper classification results, it already outperforms the first baseline ever published on this dataset by more then 4 percent accuracy. As a next step, we use the test-exemplar-specific set for patch discovery to learn what makes samples in this set distinctive, and finally apply the discovered representations to classify the query.

**Standard vs. Local Learning of Representations and Models** Classification results for standard and local approaches on both datasets are given in Fig. 2 (magenta and green line). Although only a fraction of dimensions is used (see Table 2 for an overview of numbers of patches discovered), the local learning approach outperforms the single model by $\sim 6\%$ accuracy on the small dataset. On the larger dataset, however,

[1] Source code available at http://www.inf-cv.uni-jena.de/fine_grained_recognition .

Table 1: Fine-grained recognition results on the CUB-2011 dataset.

| Approach | CUB-2011-14 | CUB-2011-200 |
|---|---|---|
| Wah *et al.* [11] | – | $10.25\%$ |
| Single representation | $58.01\%$ | $39.35\%$ |
| *Local representation ($K = 150$)* | *$63.89\%$* | *$34.16\%$* |
| Style-awareness [7] | - | $38.31\%$ |
| POOF [1] | $70.10\%$ | $56.78\%$ |
| Goering *et al.* [5] | $73.39\%$ | $57.99\%$ |
| Local representation ($K = 150$) + [5] | $\mathbf{76.64}\%$ | $\mathbf{58.55}\%$ |

Table 2: Number of discovered patch detectors for standard and local learning.

| | **CUB-2011-14** | | | | | **CUB-2011-200** | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | standard | $k = 20$ | 40 | 80 | 100 | standard | $k = 20$ | 40 | 80 | 100 |
| **# detectors** | 2,936 | 142 | 285 | 571 | 715 | 40,659 | 133 | 267 | 535 | 669 |
| **Ratio** | 1 | 4.8% | 9.7% | 19.4% | 24.3% | 1 | 0.3% | 0.6% | 1.3% | 1.6% |

a single model and representation learned on all data still results in superior performance compared to local learning results emphasized here. Nonetheless, given the fact that dimensionality is only around $2\%$ compared to the standard approach, we believe the results are promising. We conclude here that discovered representations for local approaches are compact and at the same time informative.

**Combining Patch Discovery and Part Transfer** In a final experiment, we combined the proposed scheme for local patch discovery with the semantic part transfer approach of [5] using a simple weighted combination of class probabilities. Results are given in the lower part of Table 1[2] and we conclude that state-of-the-art can be outperformed by combining the complementary techniques.

## 3 Conclusions

In this paper, we proposed learning of mid-level representations and classification models in an exemplar-specific manner. Our experimental results show that the resulting dimensionality of discovered representations is orders of magnitudes smaller compared to a single fixed representation learned from all data. At the same time, the resulting classification performance is comparable or even superior to the baseline obtained with a single model, and further combining them with semantic parts even outperforms state-of-the-art results on CUB-2011. Thus, we conclude that local learning of mid-level representations leads to compact and informative models valuable for the challenging task of fine-grained recognition.

[2] Note that in [5], reported results are overall recognition rates averaged over all samples, whereas we report average recognition rates here, *i.e.*, averaged over class accuracies.

## References

1. Berg, T., Belhumeur, P.N.: Poof: Part-based one-vs-one features for fine-grained categorization, face verification, and attribute estimation. In: Conference on Computer Vision and Pattern Recognition (CVPR). pp. 955 – 962 (2013)
2. Bottou, L., Vapnik, V.: Local learning algorithms. Neural computation 4(6), 888–900 (1992)
3. Duan, K., Parikh, D., Crandall, D., Grauman, K.: Discovering localized attributes for fine-grained recognition. In: Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3474–3481 (2012)
4. Freytag, A., Rodner, E., Darrell, T., Denzler, J.: Exemplar-specific patch features for fine-grained recognition. In: German Conference on Pattern Recognition (GCPR) (2014)
5. Göring, C., Rodner, E., Freytag, A., Denzler, J.: Nonparametric part transfer for fine-grained recognition. In: Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1–8 (2014)
6. Ionescu, R., Popescu, M., Grozea, C.: Local learning to improve bag of visual words model for facial expression recognition. In: International Conference on Machine Learning - Workshop on Representation Learning (ICML-WS) (2013)
7. Lee, Y.J., Efros, A.A., Hebert, M.: Style-aware mid-level representation for discovering visual connections in space and time. In: International Conference on Computer Vision (ICCV). pp. 1857–1864 (2013)
8. Pentina, A., Lampert, C.: A pac-bayesian bound for lifelong learning. In: International Conference on Machine Learning (ICML) (2014)
9. Singh, S., Gupta, A., Efros, A.: Unsupervised discovery of mid-level discriminative patches. In: European Conference on Computer Vision (ECCV). pp. 73–86 (2012)
10. Van De Weijer, J., Schmid, C.: Applying color names to image description. In: International Conference on Image Processing (ICIP). vol. 3, pp. III–493 (2007)
11. 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)
12. Zhang, H., Berg, A.C., Maire, M., Malik, J.: Svm-knn: Discriminative nearest neighbor classification for visual category recognition. In: Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2126–2136 (2006)