Bird Species Categorization Using Pose Normalized Deep Convolutional Nets

Steve Branson[1]
sbranson@caltech.edu

Grant Van Horn[2]
gvanhorn@ucsd.edu

Serge Belongie[3]
tech.cornell.edu

Pietro Perona[1]
vision.caltech.edu

[1] California Institute of Technology Pasadena, CA, USA

[2] University of California, San Diego
La Jolla, CA, USA

[3] Cornell Tech
New York, NY, USA

In this work we propose an architecture for fine-grained visual categorization that approaches expert human performance in the classification of bird species. We perform a detailed investigation of state-of-the-art deep convolutional feature implementations and fine-tuning feature learning for fine-grained classification. We observe that a model that integrates lower-level feature layers with pose-normalized extraction routines and higher-level feature layers with unaligned image features works best. Our experiments advance state-of-the-art performance on bird species recognition, with a large improvement of correct classification rates over previous methods (75% vs. 55-65%).

Our architecture can be organized into 4 components: keypoint detection, region alignment, feature extraction, and classification. We predict 2D locations and visibility of 13 semantic part keypoints of the birds using the DPM implementation from [1] . These keypoints are then used to warp the bird to a normalized, prototype representation. To determine the prototype representations, we propose a novel graph-based clustering algorithm for learning a compact pose normalization space. Features, including HOG, Fisher-encoded SIFT, and outputs of layers from a CNN [3], are extracted (and in some cases combined) from the warped region. The final feature vectors are then classified using an SVM.

Although we believe our methods will generalize to other fine-grained datasets, we forgo experiments on other datasets in favor of performing more extensive empirical studies and analysis of the most important factors to achieving good performance on CUB-200-2011. Specifically, we analyze the effect of different types of features, alignment models, and CNN learning methods. We believe that the results will be informative to researchers who work on object recognition in general.

Our fully automatic approach achieves a classification accuracy of 75.7%, a 30% reduction in error from the highest performing (to our knowledge) existing method [2]. We note that our method does not assume ground truth object bounding boxes are provided at test time (unlike many/most methods). If we assume ground truth part locations are provided at test time, accuracy is boosted to 85.4%. These results were obtained using prototype learning using a similarity warping function computed using 5 keypoints per region, CNN fine-tuning, and concatenating features from all layers of the CNN for each region. The major factors that explain performance trends and improvements are:

1. Choice of features caused the most significant jumps in performance. The earliest methods that used bag-of-words features achieved performance in the $10-30\%$ range. Recently methods that employed more modern features like POOF, Fisher-encoded SIFT and color descriptors, and Kernel Descriptors (KDES) significantly boosted performance into the $50-62\%$ range. CNN features have helped yield a second major jump in performance to $65-76\%$. See Figure 1.
2. Incorporating a stronger localization/alignment model is also important. Among alignment models, a similarity transformation model fairly significantly outperformed a simpler translation-based model. Using more keypoints to estimate warpings and learning pose regions yielded minor improvements in performance. See Figure 2.
3. When using CNN features, fine-tuning the weights of the network and extracting features from mid-level layers yielded substantial improvements in performance. See Figure 3.

[1] Steve Branson, Oscar Beijbom, and Serge Belongie. Efficient large-scale structured learning. In *CVPR*, 2013.

[2] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. *arXiv preprint arXiv:1310.1531*, 2013.

[3] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

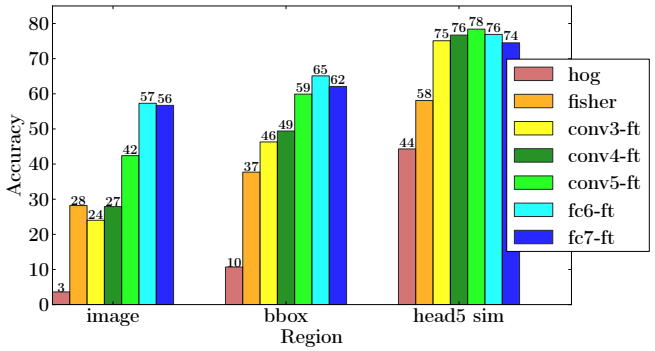

Figure 1: **Feature Performance Comparison:** CNN features significantly outperform HOG and Fisher features for all levels of alignment (image, bounding box, head).

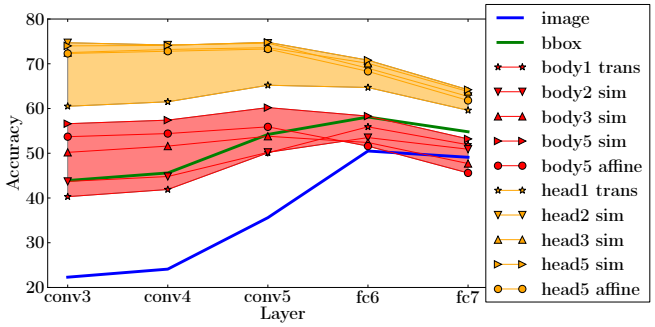

Figure 2: **Effect of CNN Layers For Different Regions:** The later fully connected layers (fc6 & fc7) significantly outperform earlier layers when a crude alignment model is used (image-level alignment), whereas convolutional layers (conv5) begin to dominate performance as we move to a stronger alignment model (from image $\rightarrow$ bbox $\rightarrow$ body $\rightarrow$ head).

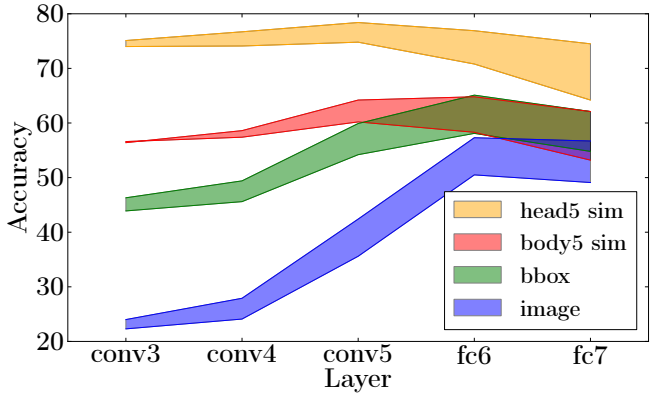

Figure 3: **Effect of Fine-Tuning with GT Parts:** Fine-tuning significantly improves performance for all alignment levels (width of each tube). Improvements occur for all CNN layers; however, the effect is largest for fully connected layers.