Part Localization using Multi-Proposal Consensus for Fine-Grained Categorization

Kevin J. Shih
kjshih2@illinois.edu
Arun Mallya
amallya2@illinois.edu
Saurabh Singh
ss1@illinois.edu
Derek Hoiem
dhoiem@illinois.edu

University of Illinois
Urbana-Champaign
IL, US

The most common approach to keypoint localization is to learn a set of keypoints detectors to model appearance and an associated spatial model [3, 4, 5, 9] to capture their spatial relations. Individual keypoint detectors typically model local appearance and thus rely on expressive spatial models to capture long range dependencies. Alternatively, the keypoint detectors could condition their predictions on larger spatial support and jointly predict several keypoints [2], then the need for expressive spatial models could be eliminated, leading to simpler models.

For effective fine-grained category detection, the keypoint localization methods must have high accuracy, low false positive rates, and low false negative rates. Missed or poorly localized predictions make it impossible to extract the relevant features for the task at hand. If a keypoint is falsely determined to be present within a region, it is hard to guarantee that it will appear at a reasonable location. In the case of localizing keypoint-defined regions of an image, such as head or torso of a bird, a single outlier in the keypoint predictions can significantly distort the predicted area. This specific case is noteworthy, as several of the current best-performing methods on the CUB 200-2011 birds dataset [8] rely on deep-network based features extracted from localized part regions [1, 3, 4, 9].

In this work, we tackle the problem of learning a keypoint localization model that relies on larger spatial support to jointly localize several keypoints and predict their respective visibilities. Leveraging recent developments in Convolutional Neural Networks (CNNs), we introduce a framework that outperforms the state-of-the-art on the CUB dataset. Further, while CNN-based methods suffer from a loss of image resolution due to the fixed-sized inputs of the networks, we introduce a simple sampling with outlier rejection scheme that allows us to work around the issue without the need to train cascades of coarse-to-fine localization networks [6, 7]. Finally, we test our predicted keypoints on the fine-grained recognition task. Our keypoint predictions are able to significantly boost the performance of current top-performing methods on the CUB dataset.

We design our model to simultaneously predict keypoint locations and their visibilities for a given image patch. Given N keypoints of interest, we train a network to output an N dimensional vector $\hat{v}$ and a $2N$ dimensional vector $\hat{l}$ corresponding to the visibility and location estimates of each of the keypoints k_i, $i \in \{1, N\}$, respectively. The corresponding groundtruth targets during training are v and l. We define v to consist of indicator variables $v_i \in \{0, 1\}$ such that $v_i = 1$ if keypoint k_i is visible in the given Edge Box image before padding is performed, and 0 otherwise. The groundtruth location vector l is of length $2N$ and consists of pairs (l_{x_i}, l_{y_i}) which are the normalized $(\tilde{x}, \tilde{y})$ coordinates of keypoint k_i with respect to the un-padded Edge Box image. Output predicted from the network, $\hat{v}_i \in [0, 1]$, acts as a measure of confidence of keypoint visibility.

To share the information across categories, our model is trained in a category agnostic manner. At test time, we efficiently sample each image with Edge Boxes, make predictions from each Edge Box, and reach a consensus by thresholding for visibility and reporting the medoid. Our method is illustrated in Fig. 1.

Results We evaluate our prediction model on the Caltech-UCSD Birds dataset [8]. This dataset contains 200 bird categories with 15 keypoint location and visibility labels for each of the total of 11788 images. We first evaluate our keypoint localization and visibility predictions against other top-performing methods and demonstrate state-of-the-art resutls in both keypoint localization and visibility. Next, we demonstrate their effectiveness in the fine-grained categorization task by using the predicted keypoints to align head and torso regions, then extracting finetuned AlexNet features from the localized regions to classify with a linear SVM. While this essentially re-creates the classification step of Zhang et al. [9], substituting in our better localized parts improves their accuracy by over 4% when ground truth bounding boxes are not provided.

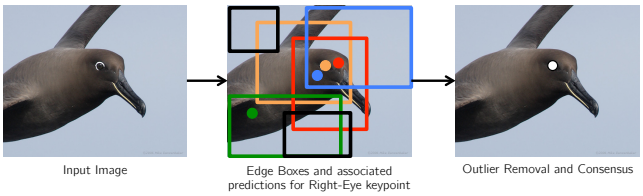

Figure 1: The pipeline of our keypoint localization process: Given an input image, we extract multiple Edge Boxes. Using each edge box, we make predictions for the location of each of the 15 keypoints, along with their visibility confidences. We then find the best predicted location by performing confidence thresholding and finding the medoid. The process is illustrated for the right eye keypoint (Black edge boxes without associated dots make predictions with confidences below the set threshold, and green is an outlier with a high confidence score).

[1] Thomas Berg and Peter N Belhumeur. Poof: Part-based one-vs.-one features for fine-grained categorization, face verification, and attribute estimation. In *CVPR*, 2013.

[2] L. Bourdev and J. Malik. Poselets: Body part detectors trained using 3d human pose annotations. In *ICCV*, 2009.

[3] Steve Branson, Grant Van Horn, Serge Belongie, and Pietro Perona. Bird species categorization using pose normalized deep convolutional nets. In *BMVC*, 2014.

[4] Jiongxin Liu and Peter N Belhumeur. Bird part localization using exemplar-based models with enforced pose and subcategory consistency. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 2520–2527. IEEE, 2013.

[5] Jiongxin Liu, Yinxiao Li, and Peter N Belhumeur. Part-pair representation for part localization. In *Computer Vision–ECCV 2014*, pages 456–471. Springer, 2014.

[6] Yi Sun, Xiaogang Wang, and Xiaoou Tang. Deep convolutional network cascade for facial point detection. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 3476–3483. IEEE, 2013.

[7] Alexander Toshev and Christian Szegedy. Deeppose: Human pose estimation via deep neural networks. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 1653–1660. IEEE, 2014.

[8] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology.

[9] Ning Zhang, Jeff Donahue, Ross Girshick, and Trevor Darrell. Part-based r-cnns for fine-grained category detection. In *Computer Vision–ECCV 2014*, pages 834–849. Springer, 2014.