WHoG: A Weighted HoG-Based Scheme For The Detection Of Birds And Identification Of Their Poses In Natural Environments

Debajyoti Karmaker*, Ingo Schiffner*, Reuben Strydom*†and Mandyam V. Srinivasan*†
* The Queensland Brain Institute, University of Queensland,St Lucia, QLD, Australia
†The School of Information Technology and Electrical Engineering, University of Queensland, St Lucia, QLD, Australia
Email: {d.karmaker,i.schiffner,r.strydom,m.srinivasan}@uq.edu.au

***Abstract*—We describe a technique for object detection that uses a combination of global shape descriptors and local point descriptors. Our system is able to represent pose using a global shape descriptor, rather than the commonly used part based representation. This approach considerably reduces computational complexity and achieves a significant performance improvement on an extensive dataset: CUB-200-2011 [31]. Our methodology is valuable for the detection of textured objects that are viewed against background clutter and possess a high degree of articulation and variation of pose, as for example in birds. We demonstrate how high and low frequency gradients can be separated to better deal with the presence of interfering textures or stripes within the body, which is a major problem in the detection of bird-like objects. Furthermore, detection accuracy is improved by integrating appropriately designed scale invariant color features into the algorithm.**

I. INTRODUCTION

As we advance towards comprehensive image understanding, precise recognition of individual objects is paramount. Object detection, which involves detecting the presence of a specified object in an image and specifying its location, is a fundamental problem in computer vision. However, proper understanding of a scene requires not only the detection of objects within it, but also identifying the objects that is, categorizing them into specific classes [1]. Visual object recognition is largely concerned with classifying object categories [1]. While this provides coarse distinctions between object classes (e.g. distinguishing between different classes of animals), in many current applications this is not sufficiently fine-tuned to deal adequately with intra class variation, where small variations need to be discriminated and classified. This area of research is referred to as fine-grained recognition, which has recently started to gain attention in the computer vision community [2], [3], [7], [8] to address challenging problems in object recognition, in particular the recognition and classification of difficult objects such as birds and glass bottles. Birds, for example, assume many poses (e.g. flying, perching, walking, and swimming) that present significant challenges for robust detection and classification. The primary aim of this paper is to build a robust system for detecting an object (in our case a bird) in any given image. As a by-product, the system also delivers information on the species of the bird and its pose, which are tools that we use to improve the accuracy of detecting the presence and location of a bird.

A prominent state-of-the-art in the domain of detection is the Deformable Part-based Model (DPM), a graphical model that considers spatial relations between various pre-defined parts of the object in our case, specific body parts of the bird, allowing articulation to be modeled more precisely [10]. DPM has also been used for the modeling of bird parts, however, efforts so far have yielded an average precision of only about 12% for bird detection [10]. Other similar recent studies also fail to produce satisfactory results for bird detection [11], [12], [13]. The models employed in these studies use a tree structure to reduce complexity, for instance, minimizing the spatial relation between two connected parts, which fails to capture higher order relationships between multiple parts. Furthermore, most of these models consider just two poses (e.g. left facing or right facing), which is insufficient for characterizing the highly dynamic nature of birds.

There is a wide range of standard features available for object detection in the computer vision community, such as scale-invariant feature transform (SIFT) [15], Speeded Up Robust Features (SURF) [14] and histogram of oriented gradients (HoG) [6]. Object detection gained traction with the HoG (Histogram of Gradients) descriptor [6], which yielded great success in the domain of human detection. Primarily, the HoG descriptor was built for pedestrian detection, where clothing and color on humans were unnecessary distractors that had to be disregarded, or filtered out in an appropriate way. However, as many species of birds differ greatly in plumage color and pattern, it can be advantageous to make use of these features, rather than ignore them, and use this additional information to bolster or reject a candidate location computed by the HoG approach. Unlike SURF and SIFT, HoG descriptors are not a local descriptor, but a global one. The HoG is a model based on vector space that computes similarity using Euclidian or cosine distances, which is highly suitable for machine learning algorithms. The HoG features structure the gradients within a cell into nine orientation bins, disregarding any edges or textures. The process is rather simple, and consists of only the summation of all the gradient directions in a cell without considering their magnitudes. However, because of its cell

normalization attribute, the model is not sensitive to global contrast. Global contrast sensitivity is still an important feature of a detection model for birds, as birds are often found to be in complete lighting disparity (e.g. when they are sitting on trees). Hence a robust model must be able to appropriately handle such inconsistencies.

In this work, we propose several ways to transform HoG features to improve bird detection, which we refer as WHoG. We demonstrate the power of strongly supervised pose clustering using part annotations, and demonstrate the benefits of constructing a detector using WHoG features to locate birds. A separate detector is incorporated using scale invariant color features to accompany the WHoG features. This improves the reliability of the detection process, vastly reducing the rate of false positives. Furthermore, incorporating color features significantly increases the probability of detecting a bird when it is viewed against a cluttered background.

Inspired by the work of Lui and Belhumeru [4], we form a bottom-up shape-based pose clustering algorithm. The model in [4] uses clusters derived from each part of the bird, leading to many clusters when multiple bird parts are considered. In this paper, we demonstrate the computational benefit of our pose-clustering method, which is based on a comprehensive representation of shape. As a result, we are able to reduce the number of models from 380,000 to 300. We also demonstrate the power of combining scale invariant color features with pose clustering to enhance the reliability of our bird detection approach.

Our paper makes the following contributions:

- Modification of the HoG-based approach to suit bird detection.
- Drastic reduction of the number of models, through the use of bottom-up shape-based clustering.
- The use of scale invariant color features from species clustering to improve robustness.
- Improved state-of-the-art precision for bird detection and pose identification in the CUB-200-2011 dataset, which is considered to be the most competitive dataset within the domain of problems requiring fine-grained recognition.

II. RELATED WORKS

The so-called Bag-of-words approach has generally been used to tackle fine-grained object recognition. However, this approach struggles to cope with the visual similarity that often exists between different classes, as they share many common visual words, and sometimes distinctions can be found only in a small number of parts of the object. These parts play a very significant role in object detection and fine-grained classification [10], [17], [18], [13], [19], [20]. Currently, a growing body of literature (e.g. [10], [17], [18], [13], [19], [20]) outlines the importance of part detection in object detection. The most effective and widely used model for object detection is the Deformable Parts Model (DPM) of Felzenszwalb et al. [10]. This model makes use of a root HoG filter at a low scale and part HoG filters at a higher resolution scale. Inspired by the work of Fischler and Elschlager [21], the DPM [10] also uses a mixture of components [23] to capture deformation and view point variation [23]. Even though minimum supervision is needed for training the DPM, the model uses latent parts as latent variables with a deformation cost that is beneficial, but because it provides no semantic understanding of these parts (one-to-one or one-to-many), constructing pose-normalization from latent parts is challenging. Other studies involving pose estimation [10], [13], [24], [25] and part localization [26], [4] use strong supervision in conjunction with a set of exemplars. We also use exemplars, but we consider part visibility, and, to improve efficacy, extract useful scale invariant color information, and show how this can be applied to reduce the rate of false positives in bird detection.

Non-parametric models can be used to operate on shapes to detect local landmarks, by fitting annotated points on top of each other over the image [28], [29]. Amberg and Vetters use a generic 3D-face model to limit its applicability to relatively rigid shapes [28], whereas Belhumeur and colleagues combine the output of local detectors with a set of exemplars that potentially capture all possible configurations [27]. The idea is to build a particular SVM for each training sample by using millions of negative samples. Our approach is similar in sprit, but we shorten the expensive training and testing phases by reducing the number of models through bottom-up hierarchal clustering.

Another line of work uses human input or interaction in the classification process [26], [27]. Our main aim in this paper is to investigate fully automated recognition methodologies that do not require human interaction. However, our model can be subsequently augmented with human interaction, if necessary, to further refine the results.

Analogous to our work, authors from [4], [16], [25], [32] demonstrate the importance of part detection in fine-grained recognition. Authors from [4], [16], [25] rely on DPM [32] for part detection and build on their methodologies. Among them [4], [16] use the CUB-200-2011 bird dataset to test their scheme. Our approach offers a significant improvement over these methods. Instead of learning from each part, we accomplish very simple but instrumental global matching and perform subsequent collaborative scoring with scale invariant color features.

III. PROPOSED METHOD

Our prime objective is to detect a bird and determine its location precisely in any given image. This objective is achieved by using pose information and species-dependent color information. In the following, we first demonstrate our pose clustering procedure and build a pose specific bird detector using a modified HoG descriptor (WHoG) that produces a score for each candidate pose, with the highest score representing the identified location for that particular pose. Next we go on to build a species-specific bird detector using scale invariant color features. So there are two CNNs working parallely; one taking pose template as an input and another color information. Figure 1 shows the overall architecture of our system.

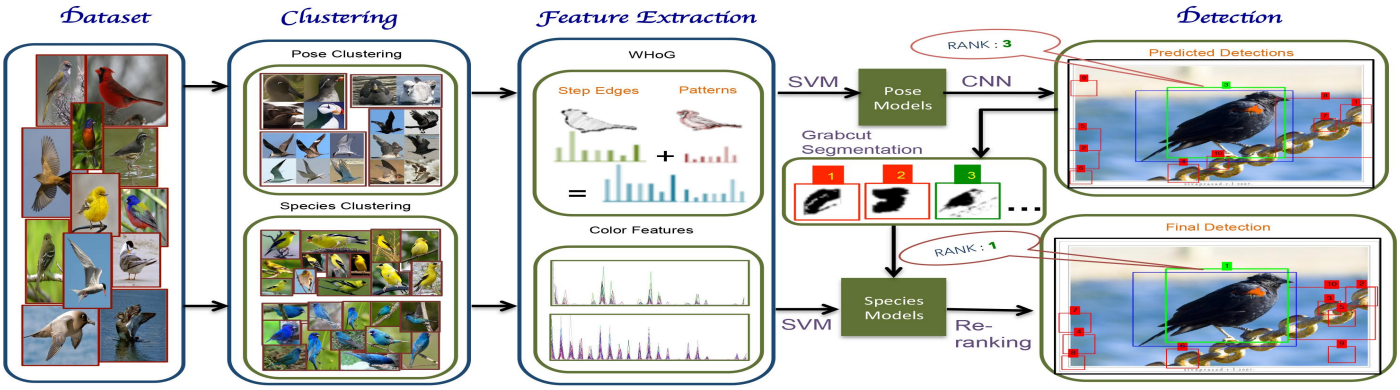

Fig. 1. Visualization of our methodology using pose and species clustering: WHoG features are extracted from the pose clusters of visually similar training images, and feed into the SVM as positive samples to generate Pose models. Scale invariant color features are extracted from the species cluster to generate a species model. Final detection is achieved by combining the outputs of the pose and species models through Grabcut segmentation. Note that in the detection phase, we only show the top 10 predictions for visual clarity, whereas for the actual experiment we used the top 20 predictions. The green box indicates the correct prediction, red boxes are the false ones and the blue box displays the ground truth. The figure is best viewed in color and by zooming in.

A. *Pose clustering*

We perform pose clustering by using part locations that capture the rough pose of a bird. Let x_i denote the i-th exemplar of a bird pose. A bounding box is set up around the image of the bird and the top left-hand corner is defined as the origin. We then represent the pose of a bird pose by a local shape matrix Δ_i^m (1) which contains the relative Euclidian distances $\Delta x_i^{j,k}$ between part locations Δx_i^k and Δx_i^j, the angles $\theta x_i^{j,k}$ between Δx_i^k and Δx_i^j, the distance $\Delta_i^{t,k}$ of each part from the origin , the angle of two corresponding parts with respect to the origin $\phi x_i^{j,k}$, and binary vectors ν_i^k that define the visibility of each part (1= visible; 0= not visible). Figure 2 shows how the 1st row is computed for the shape matrix Δ_i^m.

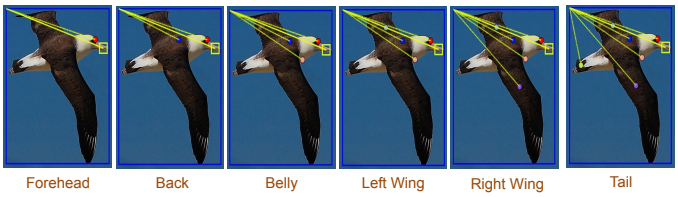

Fig. 2. Example of pose analysis. In this example the beak, marked as the yellow rectangle, is treated as the reference part to compute the relative distances between predefined neighbours (different colored circles), the angle between the reference part and the current part, the fixed distance from the origin to the current part (yellow lines), and the angles between these lines.

Fig. 3. Illustration of how other rows of the shape matrix are generated by using different parts of the birds body as a reference (marked by different color rectangles) . For each reference part the process mentioned above in Figure 2 is repeated to generate the elements of each row.

Here $\{k_1 \cdots k_{p_j}\}$ are the indices of predefined neighboring parts (p_j) containing the part locations of the beak, forehead, back, belly, left wing, right wing and tail from any given exemplar ($p_{j_m} = 7$ in our experiment). $\nu_i^k \in \{0, 1\}$ is a visibility flag. If $\nu_i^k = 0$, that means the k-th part is not visible i.e. $\Delta x_i^{j,k} = \Delta x_i^{t,k} = 0$; otherwise $\Delta x_i^{j,k}$ is computed as $x_i^k - x_i^j$, where x_i^j is the reference part and x_i^k is the current part location. $\Delta x_i^{t,k}$ is computed as $x_i^k - x_i^t$, where x_i^t indicates the origin. k_{p_j} can be any part from a predefined neighbor in a particular sequence and $k_{p_j} \neq j$. The angle between reference part and current part $\theta x_i^{j,k}$ is computed as $\cos^{-1}[\frac{\vec{j}.\vec{k}}{\|\vec{j}\|\|\vec{k}\|}]$. The angle between two corresponding parts with respect to origin $\phi x_i^{j,k}$ is computed as $\tan^{-1}[\frac{s_1 - s_2}{1 + s_1 s_2}]$. Where $s_1 = [\frac{t_2 - t_1}{k_2 - k_1}]$ and $s_2 = [\frac{t_2 - t_1}{j_2 - j_1}]$.

Figure 3 shows how the remaining rows of the shape matrix are generated. The next step is to concatenate all the rows of the matrix into a 210 element single dimensional vector. The set of all local shape vectors define the pose space, and the set of shape vectors are grouped to describe the individual pose categories. We use a bottom- up agglomerative clustering to generate T^i pose categories from pose space. Figure 4 demonstrates several examples of pose clusters generated by this procedure. Indeed, this approach is capable of clustering poses even across bird species, as long as the pose of the birds is roughly the same.

The pose detector is defined by considering all the samples from a single cluster as positive samples and treating all other cluster samples as negative samples. To increase the number of negative samples, we also take random regions of images that do not contain any part of the bird. Thus, by design, detectors are trained in such a manner as to score high on a particular pose across all bird species.

B. *WHoG (Weighted Histogram of Oriented Gradients)*

Before we go on to train the support vector machine, it is necessary to extract the boundaries of the bird (the outer edges) reliably, without being confused by the internal stripes

$$\Delta_i^m = \begin{bmatrix} \Delta x_i^{j_1,k_1} & \theta x_i^{j_1,k_1} & \Delta x_i^{t,k_1} & \phi x_i^{j_1,k_1} & \nu_i^{k_1} & \cdots & \Delta x_i^{j_1,k_{p_{j_1}}} & \theta x_i^{j_1,k_{p_{j_1}}} & \Delta x_i^{t,k_{p_{j_1}}} & \phi x_i^{j_1,k_{p_{j_1}}} & \nu_i^{k_{p_{j_1}}} \\ \vdots & & & & & \ddots & & & & & \vdots \\ \Delta x_i^{j_m,k_1} & \theta x_i^{j_m,k_1} & \Delta x_i^{t,k_1} & \phi x_i^{j_m,k_1} & \nu_i^{k_1} & \cdots & \Delta x_i^{j_m,k_{p_{j_m}}} & \theta x_i^{j_m,k_{p_{j_m}}} & \Delta x_i^{t,k_{p_{j_m}}} & \phi x_i^{j_1,k_{p_{j_m}}} & \nu_i^{k_{p_{j_m}}} \end{bmatrix} \quad (1)$$

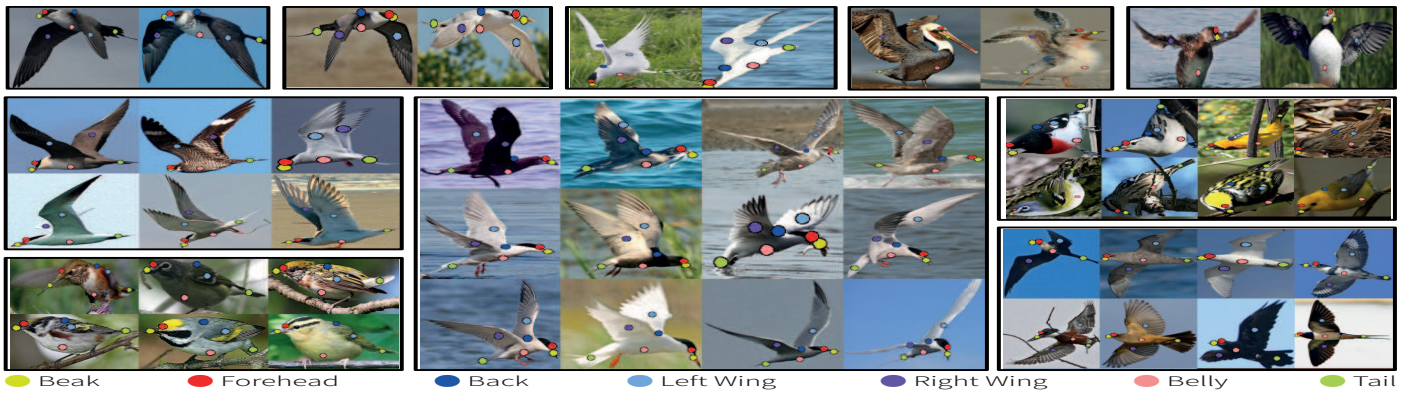

Fig. 4. Examples of identified pose clusters. Different color circles represent 7 individual body parts used to generate pose clusters. Note how the key-point arrangements are associated with the poses.

and textures as these do not represent the shape of the bird. Under natural illumination birds tend to display characteristic shading patterns on their bodies. These patterns decrease the reliability of detection, as their HoG features can be similar to those of the outer boundary. By assigning more weight to edges and less weight to body textures or stripes, we show below how WHoG can achieve the desired objective and be applicable to the detection of any object that carries internal patterns or textures.

The typical HoG algorithm initiates by computing image gradients and dividing them into 9 bins according to their orientation. Then, for every pixel, it determines the gradient image with the largest gradient magnitude at that pixel, and finally, the largest gradient magnitude is added to the histogram bin corresponding to the maximal orientation. Therefore, the HoG descriptor is sensitive only to the direction of gradients without considering the position of the gradients relative to each other, or to the distribution of directions. We separate the body stripes or texture from the boundaries by distinguishing between diffuse gradients and step edges. A step edge is an edge associated with an abrupt change in intensity, whereas a diffuse edge is one associated with a gradual change. Sometimes the textures within various body parts can be mistaken to be boundary edges because they display gradients of a similar magnitude.

We use an n-level pyramidal approach in which the image at the k-th level is denoted by $I_k(k = 1, \cdots, n)$. For each I_k, we extract the contrast α_k of the low-frequency component of the image by convolving I_k with a Gaussian low-pass filter $G_{lk}(\sigma)$ (where σ denotes the width of the kernel), and the contrast β_k of the high-frequency component by convolving I_k with a high pass filter $G_{hk}(\sigma)$, which is the spatial derivative of the low-pass filter.

We then determine the optimum pyramid level $k^{'}$ for which the quantity $\alpha_k - \lambda.\beta_k$ is a minimum.This operation determines the optimum weighting that should be applied to the low-frequency component to ensure the rejection of internal body textures. The parameter λ is chosen to optimize this operation and was set to 3.2. The final result of this procedure is an image which contains an optimally weighted combination of boundary contours and internal textures to facilitate reliable detection.

C. Species Clustering

To make color feature data well separable in RGB space, we perform a grabcut segmentation from the ground truth bounding box of each image to separate the bird from its immediate background on the basis of color differences. We then extract SIFT key points to produce scale invariant color features for each bird part. At each SIFT key point denoted by a pixel location $(i.j)$, the x and y gradients are computed separately for the red and green channel images to obtain a disparity matrix $D(i, j)$:

$$D_{i,j} = \begin{bmatrix} R_x(i,j) & G_x(i,j) \\ R_y(i,j) & G_y(i,j) \end{bmatrix} \quad (2)$$

where R_x and G_x represent gradient values along the x axis of the red and green color channels respectively. R_y and G_y denote the corresponding gradient values along the y axis. The determinant $|D_{i,j}|$ of the disparity matrix is a measure of the structural difference between the images in the red and green channels at the location (i, j), The absolute value $|D_{i,j}|$ is equal to the area enclosed by the parallelogram shown in Figure 5. The sum of the determinants ($\sum_{i,j} |D_{i,j}|$) taken over all the key points, is a measure of the overall difference between the two images in the red and green channels. This

provides a very sensitive way of discriminating between birds of different colors.

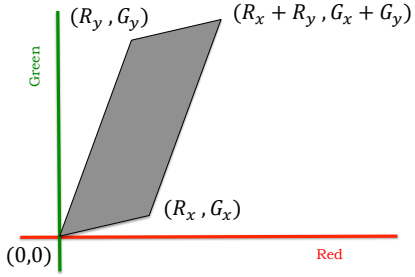

Fig. 5. Parallelogram representing image gradients the red and green channels.

The assumption is that as SIFT is rotation and scale invariant, any descriptor for these key points would also be rotation and scale invariant. We have used 5 intensity bins for each of the three-color channels, resulting in a code that can represent a total of 125 colors. To generate the color descriptor, we used gradients from only two channels - specifically, the red and green channels, because this combination displayed the greatest sensitivity to inter-species color differences in our selected data set. A histogram of the 125 colors (5x5x5) is generated by representing the area differences in terms of the red and green channel, as shown in the algorithm 1.

Algorithm 1 *Scale Invariant Color Histogram*

Input : img
Output: Hist
$img \leftarrow double(img)$ ▷ rescaling image from 0 to 1
$nBin \leftarrow 5$
$R \leftarrow img(:,:,1)$;G $\leftarrow img(:,:,2)$;B $\leftarrow img(:,:,3)$
$Rx \leftarrow R \otimes kernel$;Ry $\leftarrow R \otimes kernel'$
$Gx \leftarrow G \otimes kernel$;Gy $\leftarrow G \otimes kernel'$
$SIFTPoints \leftarrow zeros(size(img))$
$SIFTPoints[keyPointsLocations(img)] \leftarrow 1$
$Hist = zeros(nBin; nBin; nBin)$
foreach $i \leftarrow 1 : length(imgHeight)$ **do**
  foreach $j \leftarrow 1 : length(imgWidth)$ **do**
    if *SIFTPoints(i,j)* **then**
      $r \leftarrow max(1, nBin * R(i,j))$
      $g \leftarrow max(1, nBin * G(i,j))$
      $b \leftarrow max(1, nBin * B(i,j))$
      $weight \leftarrow Rx(i,j) * Gy(i,j) - Ry(i,j) * Gx(i,j)$
      $Hist(r,g,b) \leftarrow Hist(r,g,b) + weight$
    end
  end
end

Figure 6 reveals several clusters of color features where each panel represents a different species. The vertical axis in this instance represents the color intensity of the gradient image and the horizontal axes represent the total 125 bins of the color histogram. The number of species clusters is fixed at 200, as the dataset used contains sample images of birds from 200 different species. In each cluster there are roughly 30 samples and each color represents a single bird. We take particular note of the similarity between the histograms within each cluster. As for the pose clustering, another SVM detector is built for the color features. The color features of one cluster are taken as positive samples, while the color features of other clusters are considered as negative samples. Taking all positive and negative samples, we train a SVM to build our species detector. Note that the roles of the shape-based pose detector and the color-based species detector are completely different. Color features are only applied to check the outputs of the pose detector to ensure that the image of the bird in the detected window not only matches the shape but also possesses color features that match well with the color signature of one of the 200 categorized species. The combined score from the pose detector and the color detector that produces the best pose with color match is used to select the ultimate location of the bird. That is, with reference to fig. 1, the pose CNN produces a series of candidate locations for the bird 10 for each pose template, including all the scales from the image pyramid and ranks the locations according to the confidence value for each pose template. For the 300 pose models we acquire a total of 3000 candidate locations, from which we use the top 50 locations, as ranked by the confidence values provided by the CNN.

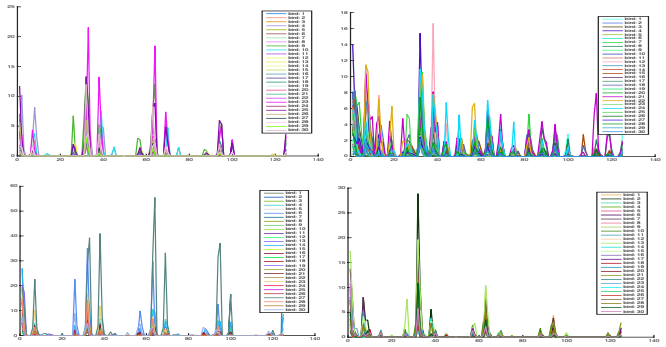

Fig. 6. Examples of scale invariant color features. Note that even though there are similar peaks, the overall histogram provides a unique identifier for each species. Note that each panel represents a different species containing roughly 30 samples. Vertical axes represent color intensity and horizontal axes represent color bins. Best viewed in color and zoomed in.

A grabcut is then performed to extract the image of the bird within the bounding box of each of the 50 candidates, and the color of the extracted image is compared with the color signatures of each of the 200 species. The color matches are given a separate ranking. The candidate location displays the highest sum of the confidence values delivered by the pose CNN and the color CNN is taken to be the highest ranked location. In the illustrated example (Figure 1), the green window in the final detection phase represents the ultimate winning location.

IV. IMPLEMENTATION DETAILS

We use SVMs implemented in VLFEAT. The features used to train the SVM for pose clustering are WHoG as described

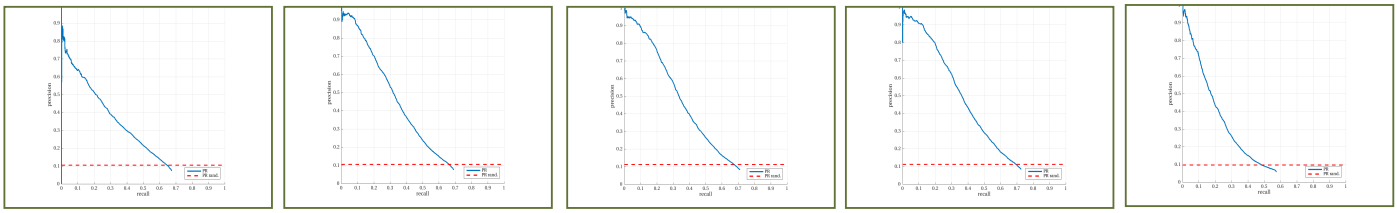

Fig. 7. Precision-recall (PR) curve shows variation of detection accuracy for cluster sizes ranging from 50 to 500. Note that the average precision(AP) is maximum for cluster size 300.

earlier. WHoG features are extracted with a cell size of 88. The scale of a bird is normalized based on other images of that cluster. To scan the image over all scales we use a scaling factor of 1.8 to build the image pyramid. We avoid an expensive sliding window approach by using Convolutional Neural Networks (CNNs) to detect birds. The pose detection and species detection are two different tasks and the features used to train them are also different. For the pose detector we build two additional WHoG descriptors, one at a finer scale and another with double the resolution of the finer scale image. Color features are extracted from a small window around each of the annotated body parts. Using grabcut segmentation, the predicted sub-windows from CNN are re-ranked with scale invariant color features to produce the final output.

V. EXPERIMENTS

We have tested our method on the CUB-200-2011 [31] dataset, which includes 11,788 images of 200 bird species (roughly 60 images per species). We used the train/test split provided in the dataset for all the experiments. Roughly 30 images per species were used for the training, and the remaining 30 for the tests. 15 different body parts were annotated with a visibility flag in each image, but only 7 parts were used to compute the shape matrix (as described earlier above). Unlike species clustering, establishing the optimum number of pose clusters to train the SVM was less clear. To determine this empirically, we tested various numbers of clusters, ranging from 50 to 500 clusters.

For evaluating the performance of our scheme, we acquired the top 20 detections based on their scores as determined by the WHoG, and calculated the precision and recall over these 20 windows. Precision and Recall (PR) were computed as $\frac{T_p}{F_p+T_p}$ and $\frac{T_p}{F_n+T_p}$ respectively, which is standard practice in object detection. T_p denotes true positives, F_p denotes false positive and F_n denotes false negatives. High precision indicates a low rate of false positives, and a high recall indicates a low rate of false negatives. High scores for both precision and recall mean that the classifier is able to produce accurate (high precision) and mostly positive results (high recall). PR curves with high recall and low precision will return many results but most of the predicted results will be incorrect. In contrast PR curves with high precision and low recall will produce few results but most of them will be correct. The average precision (AP) is thus defined as the area under the curve. Each curve was computed using a test set comprising 5794 images.

Figure 7 shows how the number of clusters affects the detection rate. We considered the top 20 detection sub-windows for each PR curve. Increasing the number of clusters reduces the sample size in each cluster, which is probably the reason why the detection rate drops for cluster sizes exceeding 300. Using only pose clustering and WHoG features we achieve the highest average precision (approximately 40%) when using 300 pose clusters. We therefore use a CNN with 300 different WHOG-based pose templates for bird detection.

For evaluating the performance of the combined approach using both pose and species clustering, we also acquired the top 20 detection sub-windows, but this time based on their combined scores. The arbitration process is that if a given location receives a rank m from the pose CNN and a rank n from the color CNN the overall rank R is a weighted sum of the two ranks: $R = K_p.m + K_c.n$, where $K_p = 0.8$ and $K_c = 1 - 0.8 = 0.2$. Again we calculated precision and recall over these 20 windows. After applying scale invariant color features average precision was 47.61% as shown in Figure 8.

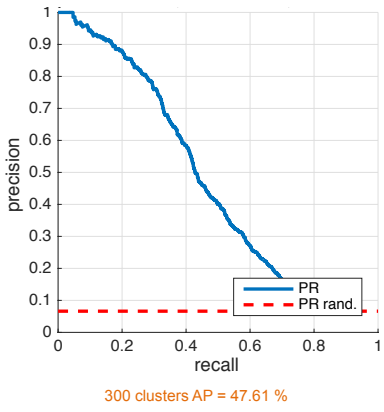

Fig. 8. Shows that the average precision improves from 39.46% to 47.61% when colour features are included.

We also compare our method with other recent detection methods for birds on the whole dataset, as well as on a subset of 14 species. Detection results, expressed in terms of average precision, are given in table 1 for the CUB-2011 data set. As can be seen, our method outperforms previous methods for the 14 species subset as well as for the whole 200 species dataset. Our method achieves much better results; due to formidable feature representation using the WHoG approach. Furthermore,

and more importantly, our implementation of species clustering using scale invariant color features has led to a significant reduction of the influence of body textures on the precision of detection. We achieve state-of-the-art performance on the dataset without using the ground truth bounding boxes or part annotations from the test data.

TABLE I
COMPARISON OF RESULTS FROM DIFFERENT METHODS. OUR METHOD SIGNIFICANTLY OUTPERFORMS ALL FOUR OF THE RECENTLY PUBLISHED TECHNIQUES. [16] AND [17] ARE NOT DIRECTLY COMPARABLE AS THEY USED AN OLDER DATASET COMPRISING ONLY 14 SPECIES.

Method	200 species	14 species
Birdlets [16]	-	40.25%
Template bagging [17]	-	44.73%
Pose pooling kernel [33]	28.18%	57.44%
Part location [4]	44.13%	62.42%
Our method	47.61%	67.42%

VI. CONCLUSION

In this paper we tackle the problem of fine-grained recognition of birds, which is a challenging problem due to the high diversity of poses that a bird can assume, and the high variation of plumage color and texture. We generate a major improvement in precision over well-known detection techniques like DPMs, which suffer from high pose variation. We provide a strategy to deal with extreme pose variations which uses bottom-up pose clustering. We also introduce and demonstrate the value of a new global feature descriptor (WHoG) that improves upon the currently used HoG descriptor. Our study demonstrates that the WHoG descriptor is well suited for object detection with diverse textures. Furthermore, by combining pose clustering and scale invariant color features i.e. combining global features with local features, we construct a powerful detector that reduces the influence of background clutter and internal body textures or stripes. In essence, this method could be utilized for more than just bird detection, and can be applied to the detection of any highly articulated object . In future work, we aim to employ techniques similar to those currently used in DPMs to further improve detection and accurate localization of individual body parts. As our model inherently contains descriptors for specific poses, we aim to use this prior knowledge to achieve detection of individual part locations in our bird model.

ACKNOWLEDGMENT

This work was partially funded by a Boeing Research Scholarship, and an ARC Distinguished Outstanding Researcher Award (DP 140100914).

REFERENCES

[1] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. IJCV, 88(2):303–338, June 2010

[2] B. Yao, A. Khosla, and F.-F. Li. Combining randomization and discrimination for fine-grained image categorization. In CVPR, pages 1577–1584, 2011.

[3] O. M. Parkhi, A. Vedaldi, C. V. Jawahar, and A. Zisserman. Cats and dogs. In CVPR, 2012.

[4] J. Liu and P. N. Belhumeur. Bird Part Localization Using Exemplar-Based Models with Enforced Pose and Subcategory Consistency. ICCV , pages 2520–2527, 2013.

[5] T. Berg and P. N. Belhumeur, POOF: Part-based one-vs.-one features for fine-grained categorization, face verification, and attribute estimation, CVPR, Jun. 2013, pp. 955–962.

[6] N. Dalal and B. Triggs, Histograms of oriented gradients for human detection, in Proc. CVPR, Jun. 2005, pp. 886–893.

[7] N. Zhang, R. Farrell, F. Iandola, and T. Darrell, Deformable part descriptors for fine-grained recognition and attribute prediction, in Proc. ICCV, Dec. 2013, pp. 729–736.

[8] C. Goering, E. Rodner, A. Freytag, and J. Denzler, Nonparametric part transfer for fine-grained recognition, in Proc. CVPR, Jun. 2014, pp. 2489–2496.

[9] T. Malisiewicz, A. Gupta, and A. A. Efros. Ensemble of exemplar-svms for object detection and beyond. Proc. ICCV, 2011.

[10] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. IEEE Transactions on Pattern Analysis and Machine Intelligence, 32(9):1627–1645, 2010.

[11] L. Bourdev, S. Maji, T. Brox, and J. Malik. Detecting people using mutually consistent poselet activations. Proc. ECCV, 2010.

[12] R. Farrell, O. Oza, N. Zhang, V. I. Morariu, T. Darrell, and L. S. Davis. Birdlets: Subordinate categorization using volumetric primitives and pose-normalized appearance. Proc.ICCV, 2011.

[13] H. Azizpour and I. Laptev. Object detection using strongly supervised deformable part models. Proc. ECCV, 2012.

[14] H. Bay, T. Tuytelaars, and L. Van Gool. Surf: Speeded up robust features. In Proc. ECCV, 2006.

[15] D. Lowe. Object recognition from local scale-invariant features. In Proc. ICCV, 1999.

[16] R. Farrell, O. Oza, N. Zhang, V. I. Morariu, T. Darrell, and L. S. Davis. Birdlets: Subordinate Categorization using Volumetric Primitives and Pose-normalized Appearance. In ICCV, 2011.

[17] B. Yao, A. Khosla, and L. Fei-Fei. Combining Randomization and Discrimination for Fine-grained Image Categorization. In CVPR, 2011.

[18] S. Yang, L. Bo, J. Wang, and L. Shapiro. Unsupervised Template Learning for Fine-Grained Object Recognition. In NIPS, 2012.

[19] Y. Yang and D. Ramanan. Articulated Pose Estimation using Flexible Mixtures of Parts. In CVPR, 2011.

[20] L. Zhu, Y. Chen, A. Yuille, and W. Freeman. Latent hierarchical structural learning for object detection. Proc. CVPR, 2010

[21] M. A. Fischler and R. A. Elschlager. The representation and matching of pictorial structures. IEEE Transactions on Computers, 22(1):67–92, January 1973.

[22] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. Proc. CVPR, 2012

[23] Y. Yang and D. Ramanan. Articulated pose estimation with flexible mixtures of parts. Proc. CVPR, 2011

[24] M. Sun and S. Savarese. Articulated Part-based Model for Joint Object Detection and Pose Estimation. In ICCV, 2011.

[25] P. N. Belhumeur, D. W. Jacobs, D. J. Kriegman, and N. Kumar. Localizing parts of faces using a consensus of exemplars. PAMI, 35(12):2930–2940, 2013

[26] S. Branson, P. Perona, and S. Belongie. Strong Supervision From Weak Annotation: Interactive Training of Deformable Part Models. In ICCV, 2011.

[27] P. N. Belhumeur, D. W. Jacobs, D. J. Kriegman, and N. Kumar. Localizing parts of faces using a consensus of exemplars. Proc. CVPR, 2011.

[28] B. Amberg and T. Vetters. Optimal landmark detection using shape models and branch and bound. Proc. ICCV, 2011.

[29] C. Wah, S. Branson, P. Perona, and S. Belongie. Multiclass recognition and part localization with humans in the loop. In ICCV, 2011.

[30] L. Bourdev, S. Maji, T. Brox, and J. Malik. Detecting people using mutually consistent poselet activations. Proc. ECCV, 2010.

[31] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The caltech-ucsd birds–200–2011 dataset. Computation and Neural Systems Technical Report, CNS–TR–2011–001, 2011.

[32] E. Gavves, B. Fernando, C. Snoek, A. Smeulders, and T. Tuytelaars. Fine-grained categorization by alignments. In ICCV, 2013.

[33] N. Zhang, R. Farrell, and T. Darrell. Pose pooling kernels for subcategory recognition. Proc. CVPR, 2012.