DETECTION OF SMALL BIRDS IN LARGE IMAGES BY COMBINING A DEEP DETECTOR WITH SEMANTIC SEGMENTATION

Akito Takeki, Tu Tuan Trinh, Ryota Yoshihashi, Rei Kawakami, Makoto Iida and Takeshi Naemura

The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-8656, Japan

ABSTRACT

This paper tackles the problem of bird detection in large landscape images for applications in the wind energy industry. While significant progress in image recognition has been made by deep convolutional neural networks (CNNs), small object detection remains a problem. To solve it, we follow the idea that a detector can be tuned to small objects of interest and semantic segmentation methods can be complementarily used to recognize large background areas. Specifically, we train a CNN-based detector, fully convolutional networks, and a superpixel-based semantic segmentation method. The results of the three methods are combined by using support vector machines to achieve high detection performance. Experimental results on a bird image dataset show the high precision and effectiveness of the proposed method.

Index Terms— object detection, semantic segmentation, CNN, FCN, birds

1. INTRODUCTION

Wind turbines used for the generation of energy are considered serious threats to endangered bird species [1], and operators now have to make assessments of bird habitats around planned sites [2]. Automatic bird detection has hence drawn the attention of industry, as it can reduce the cost and increase the accuracy of investigations in comparison with manually conducted surveys, and it may also assist automatic systems that decelerate the blades or sound an alarm at the approach of birds.

In such an application, problems that have different characteristics from general object detection occur. The shape and color of birds are clearly represented in recent image recognition datasets [3, 4, 5]. However, in data gathered by wide-area surveillance cameras, bird images are at low resolution compared with the entire image. Consequently, both the colors and shapes may be so vague that the birds cannot be detected. In addition, the background of the area under surveillance including the wind turbines, sky, clouds, sea, and forests may confuse the detector, leading to misdetections. Moreover, as a result of using a fixed camera, birds appear relatively less frequently; thus, these misdetections should be reduced as much as possible. Finding small objects in large background images has so far proven to be a difficult problem for general image recognition methods because of the large differences in resolution [6, 7].

To solve these problems, this paper proposes a method that detects small birds in large landscape images. Following the previous approaches (*e.g.*, [8]), a detector can be tuned to small objects of interest, and larger areas can be recognized by using other methods such as semantic segmentation. Because of its success in many image recognition tasks, we decided to use a successor [9] of convolutional neural networks (CNNs) [10, 11] for small bird detection. For larger areas, we use fully convolutional networks (FCNs) [12] and SuperParsing [13], wherein the former has the advantage of simultaneously detecting birds and recognizing the background, while the latter is more suited for background recognition. Linear SVMs [14] are used to combine all of the detection results.

The proposed method was experimentally evaluated with a bird dataset especially constructed for ecological investigations around wind farms. We show that the detector-based method and semantic-segmentation-based methods complement each other well; together, they yielded significantly high precision in the bird detection task.

Related work Before the recent remarkable progress of convolutional neural networks (CNNs), handcrafted features were often used as feature extractors, and they were combined with feature embeddings methods. Detection and classification could be done with the the represented features by using classifier methods, such as boosting [15], SVMs [14], and random forests [16]. An epoch-making change has occurred with the advances in CNNs and the growing availability of large-scale image datasets. Stronger learning models [17, 18] as well as more effective techniques for suppressing overfitting [19] and avoiding the vanishing gradient problem [20] have significantly improved the performance of CNNs.

Along with the advances in CNNs, many new detection methods have been proposed, where the main focus is in how to perform accurate region proposals and how to speed up the process. In region-based CNN methods (R-CNN) [21], a selective search [22] is first used to identify potentially salient object regions (called region proposal), from which image features are extracted by CNNs and classified by SVMs. We utilize ResNet [9], one of the most successful networks in

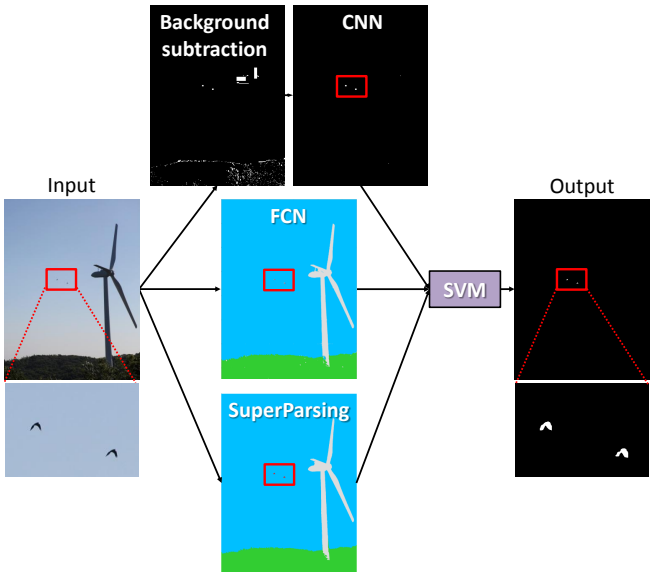

Fig. 1. Overview of the proposed method

detection, while we leave the region proposals as future work and use background subtraction for candidate region selection in this study.

Tremendous progress has also been made in semantic segmentation. There has been much debate about how to parse both background categories (*stuff*), which account for larger parts of images, and object categories (*things*), which account for smaller parts of images. Various methods parse stuff and things separately with region-based and detector-based methods [8, 23, 24].

Recently, numerous semantic segmentation methods have been proposed that are based on FCNs [12, 25, 26]. In particular, FCNs can obtain a coarse object-label map from the networks by combining the final prediction layer with lower layers (skip-layer) [27, 28, 29], where the context and localization information are available for pixel-wise labeling. This paper shows FCNs can be a complement to both a detector of things and a parser of stuffs, and together they yield high performance.

Also in the context of object detection, some methods [30, 31] use semantic segmentation methods such as FCNs. Inside-outside net [31] constructs networks using skip-layers, which simultaneously perform region proposals and classification, and improves small object detection. However, small-object detection has so far been harder than normal-size or large object detection. The smaller the object is, the lower the detection accuracy becomes [7].

2. METHOD

An overview of the proposed method is illustrated in Fig. 1. An input image is fed into three pipelines: (1) ResNet-based CNNs as a deep-feature-based detector with a background subtraction preprocess; (2) FCNs as a method that works as a detector but also as a semantic segmentation, both of which utilize deep features; and (3) SuperParsing as a hand-crafted-feature-based semantic segmentation. The class likelihoods and scores derived from them are combined using SVMs. The method outputs regions judged to be images of birds.

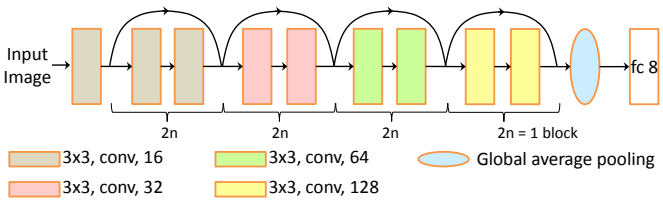

Fig. 2. Example CNN architecture for small bird image detection

2.1. CNNs for bird detection

We designed the CNNs based on ResNet [9], which achieved the best results in the detection and classification of ILSVRC 2015. In ResNet, the input of a convolutional (conv) layer bypasses one or more layers and is added to the outputs of the stacked layers. Compared with previous net structures, ResNet has fewer parameters. Moreover, even with deeper structures, gradients will not easily explode or diverge.

Fig. 2 shows our network architecture based on ResNet. We assume the sizes of the bird images ranges from 10 to 200 pixels square; thus, we design the networks to take 64×64 images as inputs, doubled the size of the original. Any size of detected bounding box will be fitted to 64×64 and fed into the networks. Because of this, one more block (the layers in yellow) is added before the global average pooling in order to capture features effectively with more hierarchies. Experimentally, the combination of four blocks with n=2 produces the best results; four blocks with n=3 produce similar results but require a longer training time, fewer blocks have less accuracy even with larger n, and more blocks cause overfitting even with fewer n.

The rest of the networks follows [9]; here, we briefly explain it for completeness. In every conv layer, the size of the kernels is 3×3. The very first conv layer has 16 kernels. Accordingly, there are four blocks, each of which includes four ($2n$ with n=2) conv layers. The number of kernels is 16, 32, 64, and 128 in each block, respectively. The first of four conv layers in the second and later blocks includes a stride of two subsamples, and this reduces the feature map size into half. Thus, the feature map size (64×64) becomes 64, 32, 16, and 8, after the process of each respective block. Finally, the ends of convolutions are connected using global average pooling, an eight-way fully connected layer (fc 8), and softmax. We use 18 stacked weighted layers in total.

2.2. Combining class likelihoods by SVM

We modified FCNs and SuperParsing to have four classes (*i.e.*, bird, sky, forest, and wind turbine), and CNNs have eight classes from its architecture, which we selected them as follows: bird, blade, tower, anemometer, nacelle, hub, forest, and other. The implementation details of SuperParsing and FCNs are provided in the training section.

Each of the three pipelines yields a class-wise likelihood or score: SuperParsing and FCNs generate pixel-wise likelihoods of classes, whereas CNNs generate a bounding box-wise score of the likelihoods of classes. If all pixels in the images were used for training the SVMs, the amount of computations would be too large to finish within a reasonable amount of time. Consequently, we use only the pixels at the center of the bounding boxes of candidate regions proposed by the inter-frame difference method. After the training in the first round, we use hard negative mining to reduce false positives and to improve the overall performance. Specifically, image regions of anemometers, night lights, the lower parts of nacelles, in which the FCNs often produce misdetections, are added for SVM training. Furthermore, the pixels collected by the inter-frame difference have statistical difference from the true pixel distribution. Because of this, when CNNs are simply combined with semantic segmentation based methods, the whole framework tends to include many misdetections by CNNs; thus, we add the background regions (sky, cloud, forest, and wind turbine) inside the candidate bounding boxes in the SVM training.

3. EXPERIMENTAL RESULTS

We implemented CNNs, FCNs and SuperParsing, as well as AdaBoost with Haar-like feature [32, 15] as a baseline. Then, we also trained several combinations of methods with our proposed framework, and evaluated their performance using a wide-area surveillance dataset of wild birds [33].

Dataset First, we picked out 82 images with different weather conditions from the dataset, which contains a set of images with 2806×3744 pixels taken nearby a wind turbine. The images were manually annotated into four classes: bird, wind-turbine, sky, and forest.

We omitted five images that were too dark due to stormy weather, and used the remaining 77 images for training of SuperParsing and FCNs. Specifically, for FCNs, the images were cropped to 500×500 pixels, because the original images were too large to process with FCNs on our GPU memory. Cropping the entire image randomly may cause there to be frames with only the sky labels, because more than a half of each image was occupied by sky. With this in mind, we performed cropping around the wind-turbine area more intensively, and obtained 70 frames from each image by shifting a 500×500 pixel window through the area. Eventually, we had 77×70=5390 frames for training FCNs.

The training images for ResNet were acquired as candidate regions of moving objects with background subtraction. The training images include bird and non-bird regions, and we prepared a class of bird and seven background classes. These extra classes help training the networks because they are frequently included in the candidate regions and likely to cause misdetection. To train the AdaBoost with Haar-like features, we used 15705 bird images and 18688 non-bird images similarly collected to train ResNet.

Evaluation We conducted the evaluation on 44 of the 77 labeled images that included more birds (183 in total) than the others. We quantified the performances of the method by using the F-measure, *i.e.*, the harmonic mean of precision and recall. In the evaluation, we regarded detected bounding boxes that had any overlap with groud-truth boxes as correct detections and boxes with no overlap as misdetection. Similarly, in segmentation-based methods, we regarded the outputs that had any region of overlap with the ground truth as correct detections, and those without overlap as misdetections.

Training *SuperParsing:* We trained SuperParsing with the 77 images annotated with four classes by using 11-fold cross validation. In [13], to form retrieval sets semantically similar to an input image, the size of the retrieval sets were set to 200. But as we had only 77 annotated images, we retrieved all of the images.

FCNs: We used an FCN-8s model [12] pretrained on PASCAL-Context [34], which contains 59 category (+ background) segmentations. The 59 (+ 1) classes include bird and sky, but forest or wind turbine are not included. Alternatively, we utilize tree and airplane classes, as pretrained classes of forest and wind turbine, respectively. We then fine-tuned the model with the prepared images for FCNs by using two-fold cross validation.

CNNs: We trained the ResNet based model with eight-class training images from scratch. We used a weight decay of 0.0001 and momentum of 0.9, and the method described in [35] for initializating the weights (*i.e.*, in the same way as [9]). In addition, we used batch normalization [20] to reduce the internal covariate shift and accelerate learning. A batch normalization layer was added to the output of every convolutional layer.

Haar+AdaBoost: AdaBoost with Haar-like features was trained following [33]. Moving object candidates were detected by the inter-frame difference. Then, the object candidates were marked with square bounding boxes and trained the detector from the bird and non-bird labels.

SVMs: We combined the class likelihoods or scores by using pixel-wise SVM training and evaluated the performances of the individual methods and their combinations.

Results Fig. 3 shows examples of detection results on the bird image dataset intended for ecological investigations. More results can be found in the supplementary material.

Method	Precision	Recall	F-measure
HA	0.064	0.514	0.114
SP	**1.000**	0.366	0.536
FCN	0.684	0.519	0.590
CNN	0.598	**0.902**	**0.719**
SP*	0.989	0.508	0.672
FCN*	0.709	0.585	0.641
FCN+SP	**1.000**	0.546	0.707
CNN+SP	0.950	0.618	0.748
CNN+FCN	0.924	0.798	0.856
Proposed (CNN+FCN+SP)	0.955	**0.803**	**0.872**

Table 1. F-measure of various methods. * represents the method combined with SVMs.

Size	Method	Precision	Recall	F-measure
tiny	FCN+SP	**1.000**	0.030	0.058
	CNN+SP	0.826	0.284	0.422
	CNN+FCN	0.808	0.627	0.706
	CNN+FCN+SP	0.860	**0.642**	**0.735**
small	FCN+SP	**1.000**	0.800	0.889
	CNN+SP	0.969	0.775	0.861
	CNN+FCN	0.972	**0.863**	0.914
	CNN+FCN+SP	**1.000**	**0.863**	**0.926**
normal	FCN+SP	**1.000**	0.944	0.971
	CNN+SP	**1.000**	0.890	0.941
	CNN+FCN	**1.000**	**0.972**	**0.986**
	CNN+FCN+SP	**1.000**	**0.972**	**0.986**

Table 2. F-measure of various methods by size

We counted the true positives (TP) and false positives (FP) of birds and calculated the precision, recall and F-measure. The results are summarized in Table 1. AdaBoost with Haar-like features and SuperParsing are denoted as HA and SP, respectively. In addition, SP* and FCN* represent the method combined with SVMs. Usually, SP or FCNs output class label with the highest likelihood, while SVMs consider all of the class likelihoods for the output through learning.

The upper part of Table 1 shows the results of individual methods. SP achieved the highest precision, while CNNs achieved the best recall rate. FCNs achieved the intermidiate score between SP and CNNs. As expected, CNNs highly outperform HA. SP* and FCN* performed better than the ones without SVMs, showing that SVMs have the ability to maximize the potential of semantic segmentation methods.

The lower part of Table 1 shows the results of combination of methods, where all combinations exceed each single method in terms of F-measure. Particularly, combinations with SP have higher precision, suggesting that SP can suppress false positives because it can recognize backgrounds well. The CNN+FCN result shows FCNs also can recognize backgrounds. The proposed method achieved the highest F-measure.

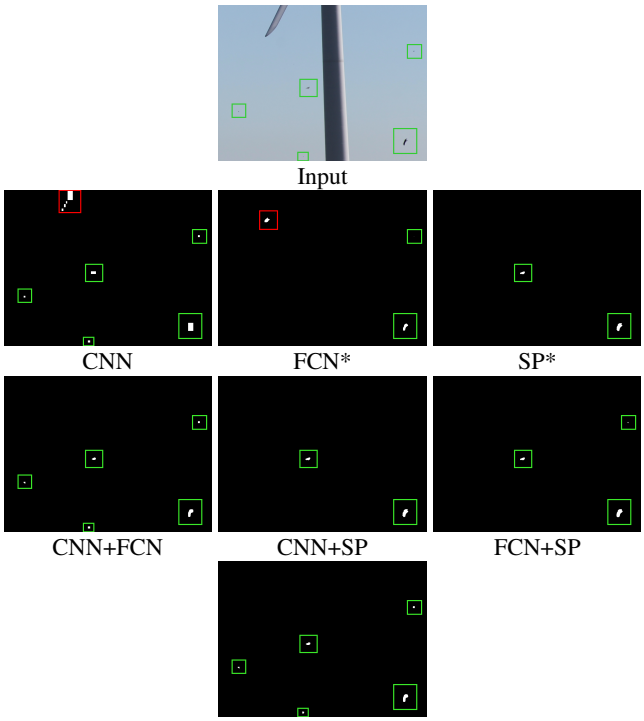

Fig. 3. Examples of detection results on the bird image dataset intended for ecological investigations. The green squares mean TP. The red squares mean FP.

To show the robustness of our method to the varying size of the bird images, Table 2 summarizes the results according to image size. The three image sizes, tiny ($\leq$15$\times$15), small ($\leq$45$\times$45), and normal ($>$45$\times$45) are determined according to [7]. In all image sizes, the proposed method produces the best F-measure. SP is not suited for detecting tiny images of birds, while when combined with both CNN and FCN, SP can work effectively. In the case of small and normal, CNN+FCN achieved the highest F-measure, followed by FCN+SP and CNN+SP.

4. CONCLUSION

We combinined a CNN-based detector with a fully convolutional network and a superpixel-based semantic segmentation by using support vector machines to achieve high performance in detecting small objects in large images. Experiments on a bird image dataset intended for ecological investigations, showed that our method detects birds with high precision. In addition, we experimentally elucidated the role of each method in small object detection.

Acknowledgment: A part of this work is entrusted by the Ministry of the Environment, JAPAN (MOEJ), the project of which is to examine effective measures for preventing birds, especially sea-eagles, from colliding with wind turbines.

5. REFERENCES

[1] K. S. Smallwood, L. Rugge, and M. L. Morrison, "Influence of behavior on bird mortality in wind energy developments," *The Journal of Wildlife Management*, 73(7):1082–1098, 2009.

[2] S. Bassi, A. Bowen, and S. Fankhauser, "The case for and against onshore wind energy in the uk," *Grantham Res. Inst. on Climate Change and Env. Policy Brief*, 2012.

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

[4] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona, "Caltech-ucsd birds 200," 2010.

[5] G. Van Horn, S. Branson, R. Farrell, S. Haber, J. Barry, P. Ipeirotis, P. Perona, and S. Belongie, "Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection," in *CVPR*, 2015.

[6] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, et al., "Imagenet large scale visual recognition challenge," *IJCV*, pp. 1–42, 2014.

[7] B. Pepik, R. Benenson, T. Ritschel, and B. Schiele, "What is holding back convnets for detection?," in *Patt. Recog.* 2015.

[8] J. Tighe and S. Lazebnik, "Finding things: Image parsing with regions and per-exemplar detectors," in *CVPR*. 2013.

[9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *arXiv preprint arXiv:1512.03385*, 2015.

[10] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, et al., "Backpropagation applied to handwritten zip code recognition," *Neural computation*, 1(4):541–551, 1989.

[11] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012.

[12] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *CVPR*, 2015.

[13] J. Tighe and S. Lazebnik, "Superparsing," *IJCV*, 101(2):329–349, 2013.

[14] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, 20(3):421–436, 1995.

[15] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of computer and system sciences*, 55(1):119–139, 1997.

[16] J. Marin, D. Vázquez, A. M. López, J. Amores, and B. Leibe, "Random forests of local experts for pedestrian detection," in *ICCV*, 2013.

[17] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *ICLR*, 2015.

[18] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *CVPR*, 2015.

[19] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *JMLR*, 15(1):1929–1958, 2014.

[20] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *ICML*, 2015.

[21] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *CVPR*, 2014.

[22] J. R. Uijlings, K. E. van de Sande, T. Gevers, and A. W. Smeulders, "Selective search for object recognition," *IJCV*, 104(2):154–171, 2013.

[23] R. Mottaghi, S. Fidler, J. Yao, R. Urtasun, and D. Parikh, "Analyzing semantic segmentation using human-machine hybrid crfs," 2013.

[24] J. Dong, Q. Chen, S. Yan, and A. Yuille, "Towards unified object detection and semantic segmentation," in *ECCV*. 2014.

[25] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected crfs," in *ICLR*, 2015.

[26] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. Torr, "Conditional random fields as recurrent neural networks," in *ICCV*, 2015.

[27] P. Sermanet, K. Kavukcuoglu, S. Chintala, and Y. LeCun, "Pedestrian detection with unsupervised multi-stage feature learning," in *CVPR*. 2013.

[28] W. Liu, A. Rabinovich, and A. C. Berg, "Parsenet: Looking wider to see better," *CoRR*, abs/1506.04579, 2015.

[29] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik, "Hypercolumns for object segmentation and fine-grained localization," in *CVPR*, 2015.

[30] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "Overfeat: Integrated recognition, localization and detection using convolutional networks," in *ICLR*. April 2014.

[31] S. Bell, C. L. Zitnick, K. Bala, and R. Girshick, "Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks," *arXiv preprint arXiv:1512.04143*, 2015.

[32] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *CVPR*. 2001.

[33] R. Yoshihashi, R. Kawakami, M. Iida, and T. Naemura, "Construction of a bird image dataset for ecological investigations," in *ICIP*. 2015.

[34] R. Mottaghi, X. Chen, X. Liu, N.-G. Cho, S.-W. Lee, S. Fidler, R. Urtasun, and A. Yuille, "The role of context for object detection and semantic segmentation in the wild," in *CVPR*. 2014.

[35] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *ICCV*, 2015.