Towards Automatic Wild Animal Detection in Low Quality Camera-trap Images Using Two-channeled Perceiving Residual Pyramid Networks

Chunbiao Zhu[1], Thomas H. Li[2] and Ge Li[1*]
[1]SECE, Shenzhen Graduate School
Peking University
Shenzhen, China
[2]Gpower Semiconductor Inc
Suzhou, China
zhuchunbiao@pku.edu.cn, thomas.li@gpower-semi.com, *geli@ece.pku.edu.cn

Abstract

Monitoring animals in the wild without disturbing them is possible using camera trapping framework, which is a technique to study wildlife using automatically triggered cameras and produces great volumes of data. However, camera trapping collects images often result in low image quality and includes a lot of false positives (images without animals), which must be detection before the post-processing step. This paper presents a two-channeled perceiving residual pyramid networks (TPRPN) for camera-trap images objection. Our TPRPN model attends to generating high-resolution and high-quality results. In order to provide enough local information, we extract depth cue from the original images and use two-channeled perceiving model as input to training our networks. Finally, the proposed three-layer residual blocks learn to merge all the information and generate full size detection results. Besides, we construct a new high-quality dataset with the help of Wildlife Thailand's Community and eMammal Organization. Experimental results on our dataset demonstrate that our method is superior to the existing object detection methods.

1. Introduction

Our wildlife population is increasingly threatened because human behavior is changing the natural system through aggressive resource acquisition and landscape changes. In addition, the urbanization of our society has reduced the interaction between humans and wildlife, and many outdoor recreation activities have decreased in popularity. As a result of this problem, our society has caused more problems for wildlife, while also reducing the focus on wildlife species and natural ecosystems. This has created a major barrier to effective management of natural resources and wildlife conservation.

Studying and monitoring wildlife can be achieved by means of non-invasive sampling techniques such as the camera trapping approach [14, 4, 8]. This method captures digital images of wild animals, using small devices composed of a digital camera and a passive infrared sensor. Camera trapping helps the biologist to sample animal populations and to observe species for conservation purposes, e.g. delineating species distributions, monitoring animal behavior, and detecting rare species [16, 19, 17, 13].

Figure 1. Examples of challenging image conditions.

Although camera trapping is a useful methodology in ecology, this method generates a large volume of images, there are many challenges in camera-trap images due to environmental conditions, animal behavior, and hardware limitation. Therefore it is a big challenge to process the recorded images and even harder, if the biologists are looking to identify all photographed species. In order to help biologists to reduce a large number of redundant work, au-

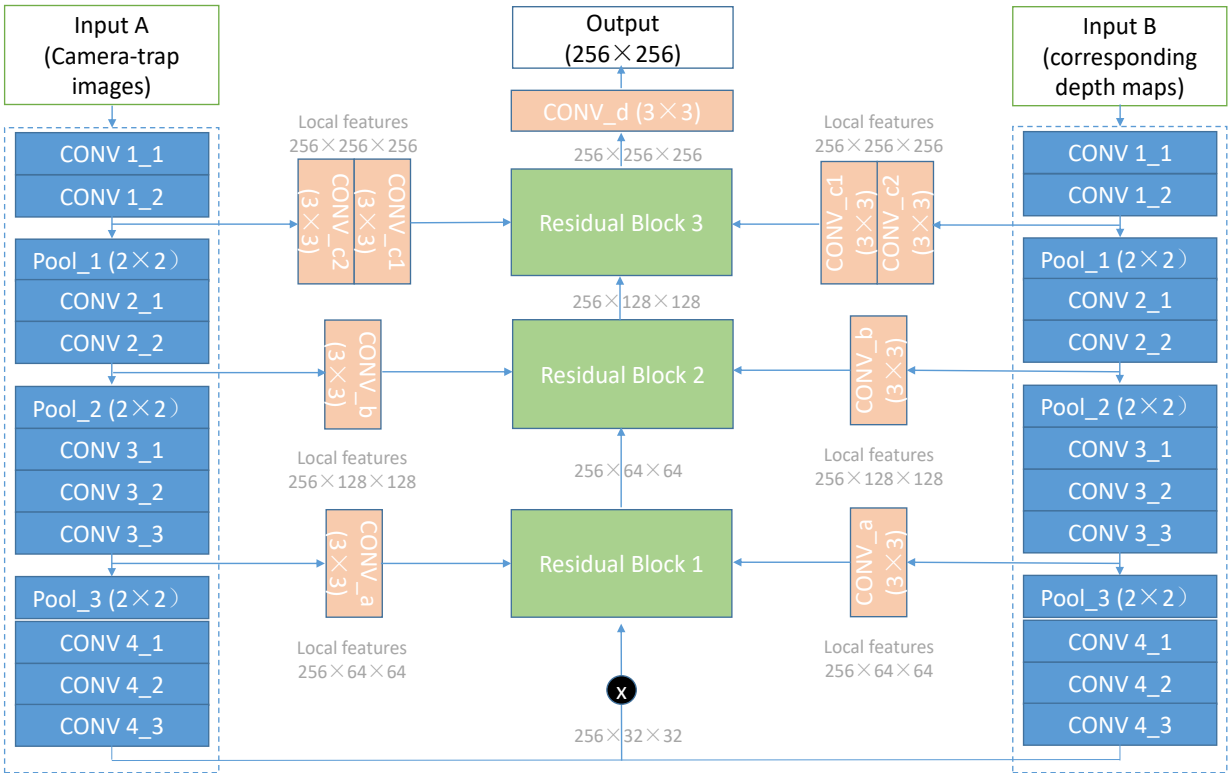

Figure 2. The framework of the two-channeled perceiving residual pyramid networks(TPRPN).

tomatic processing the camera-trap images methods should be used.

As a pre-processing stage of animal classification, automatic detection of animal in camera-trap images still remains an unsolved problem due to very challenging image conditions (Fig.1). Figure 1 shows some examples of challenging image conditions in camera-trap images.

A few previous works [15, 18, 10, 2] proposed solutions for this problem. Shukla et.al [10] design a simple pipeline comprising of superpixel segmentation, texture based feature extraction followed by mean shift clustering using the learned metric. Giraldozuluaga et.al [2] presents a Multi-Layer Robust Principal Component Analysis (RPCA) for camera-trap images segmentation, which uses histogram equalization and Gaussian filter as pre-processing, texture and color descriptors as features, and morphological filters with active contour as postprocessing.

Although, those approaches can segment most of the animals in camera-trap images, it is very difficult to produce good results when an animal has low features contrast compared to the background. In this work, we propose a two-channeled perceiving residual pyramid networks to solve the aforehand problem. First, we extract depth cue from the original images. Then, we use two-channeled perceiving model as input to training our networks. Finally, the proposed three-layer residual blocks learn to merge all the information and generate detection results.

To evaluate the proposed method, we have constructed a new dataset collected from more than 10 thousand camera-trap images. Extensive experimental evaluations show that the proposed method achieves superior performance than the existing object detection methods tested on this new dataset.

The rest of the paper is organized as follows: in section 2 the TPRPN algorithm is described. Section 3 describes the experiments and results used to test the models. Finally, in section 4 conclusions are presented.

2. Proposed Algorithm

2.1. Depth Map Generation

We collect the nearest frame of the original image as its image pairs.

Then, given an image pair, we can easily obtain a rough depth map using image correspondences. Since, most of the stereo matching algorithms may not be reliable in complex scenes, due to large depth range or flat regions, we first use Flow [12] to generate the flow map, which is robustness in both indoor and outdoor scenes.

Then we get the depth map according to the flow map.

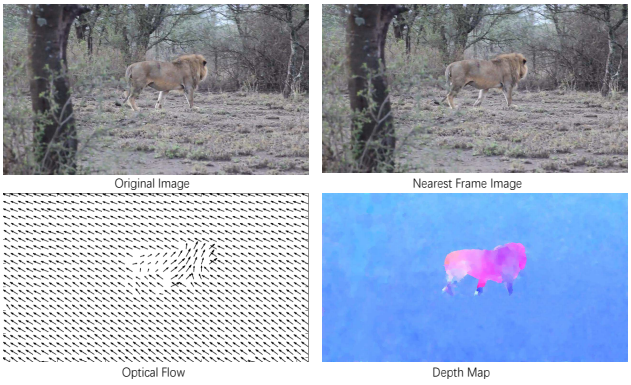

Figure 3. The visual process of depth map generation.

We get the depth map by utilizing the RGB information. Specifically, SLIC [6] is applied to over-segment the color image. As the color image and the rough depth map are aligned, we calculate the significant peaks of the depth histogram within each superpixel on the rough depth map. Suppose that a peak contains n_p pixels, the bins next to the peak contain n_l and n_r pixels respectively, and the superpixel contains N pixels. A peak is considered significant if the following conditions are satisfied:

$$\frac{n_p}{N} \geq \delta_1, min\frac{n_p}{n_l}, \frac{n_p}{n_r} \geq \delta_2, \tag{1}$$

where δ_1 and δ_2 are the threshold value.

Pixels inside the superpixel are assigned with the average depth value of the nearest peak. This process can help smooth the rough depth map, remove tiny noisy areas, fill blank holes, and refine object boundary errors (Fig.3).

2.2. Two-channeled Perceiving for Deep Features

The pre-trained VGG16 [11] networks are used in our method. We use a part of the convolutional layers and remove all the fully connected layers, so those fully connected layers are not shown in Fig.2.

In our work, the input size is $3 \times 256 \times 256$, so that both the widths and heights of the feature maps output from these layers are 32. And we use original images and corresponding depth maps as input to get both two-channeled perceiving deep features.

2.3. Three-layer Residual Blocks

As residual networks [3] achieved the state-of-the-art results on image classification, we employ the residual architecture in our deep networks. Denoting the input and output respectively as x_i and y_i, and formula of the original residual block is:

$$y_i = F(x_i) + x_i, \tag{2}$$

where F(x) is a residual function. In our three-layer residual block, we add deep local features extracted from pre-trained networks as extra inputs. Denote local features as u, the output of three-layer residual block is:

$$y_i = R_i + U_p(x_i), \tag{3}$$

$$R_i = F(x_i, u_i), \tag{4}$$

where U_p is a upsample layer. x_i are upsampled by scale 2, and x_i and μ_i are concatenated in depth. Then we use convolutional layers to learn the residual R_i.

For refinement, we aim to generate large and high quality feature maps after each step. In our three-layer residual block, the width and height of output y_i are twice over the size of x_i, and additional local information μ_i are used. Unlike the unpooling layer which only recovers limited information, we directly extract the local features before the pooling layers in the pre-trained networks. In this way, we hypothesize that all the lost information can be recovered and used to generate high quality feature maps.

Pooling operation loses local information, and we aim to recover this information for generating high-resolution and high-quality detection maps. We extract deep features from $conv_{1\_1}$, $conv_{2\_2}$ and $conv_{3\_3}$ in pre-trained networks, and before those features are input into three-layer residual blocks, they respectively pass through a batch normalization layer and one (for features from $conv_{2\_2}$ and $conv_{3\_3}$) or two (for features from $conv_{1\_2}$) convolutional layers. Local features are added into three-layer residual blocks. And after each three-layer residual block, the widths and heights are doubled. Finally, we can get a full size detection map.

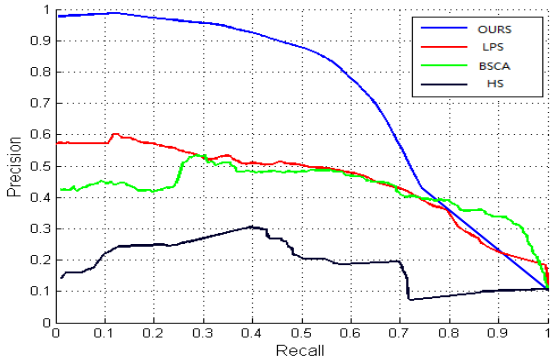

Figure 4. The PR curve evaluation result on VWM dataset.

3. Experimental Results

3.1. Datasets

We collect the VWM dataset with the help of the Wildlife Thailand, which is a community website for sharing information, photographs and experiences on Thailand's wildlife, nature and protected areas in order to help

Method	MAE	F-measure
HS	0.3294	0.2042
BSCA	0.2423	0.4179
LPS	0.1189	0.4754
TPRPN	**0.0722**	**0.7303**

Table 1. MAE and F-measure Results. Ours is better.

everyone have the opportunity to explore Thailands outstanding wildlife and National Parks. The dataset contains more than 100 camera-trap images with the manual annotation which are collected from 12 videos.

In our experiments, we use the training dataset and validation dataset from MSRA10K [1], which contains many objects in the real world. We only use our VWM dataset as the testing dataset, which can avoid impacting comparison test results.

3.2. Evaluation indicators

Experimental evaluations are based on standard measurements including precision-recall curve, MAE (Mean Absolute Error), F-measure. The precision is defined as:

$$Precision = \frac{\|p_i \mid d(p_i) \geq d_t \cap p_g\|}{\|p_i \mid d(p_i) \geq d_t\|}, \tag{5}$$

where $p_i \mid d(p_i) \geq d_t$ indicates the set that binarized from a detection map using threshold d_t. p_g is the set of pixels belonging to groundtruth object.

The recall define as:

$$Recall = \frac{\|p_i \mid d(p_i) \geq d_t \cap p_g\|}{\|p_g\|}. \tag{6}$$

The precision-recall curve is plotted by connecting the P-R scores for all thresholds.

The MAE is formulated as:

$$MAE = \frac{\sum_{i=1}^{N} \|GT_i - S_i\|}{N}. \tag{7}$$

where N is the number of the testing images, GT_i is the area of the ground truth of image i, S_i is the area of detection result of image i.

The F-measure is formulated as:

$$Fmeasure = \frac{2 \times Precision \times Recall}{Precision + Recall}. \tag{8}$$

3.3. Comparison

We compare the performance of our method with several state-of-the-art object detection methods. Including HS [9], BSCA [7] and LIP [5]. We use the codes provided by the authors to reproduce their experiments. For all the compared methods, we use the default settings suggested by the authors. We use our dataset to evaluate the performances. We randomly select some result maps of these methods and show them in Fig.5. The proposed method TPRPN can better detect animals, and our method is more similar to the ground truth.

As shown in Figs.4 and Table.1, our TPRPN has a huge improvement over previous state-of-the-art methods.

4. Conclusion

In this paper, we proposed a two-channeled perceiving residual pyramid networks towards automatic wild animal detection in low quality camera-trap images. We extract depth cue from the original images and use two-channeled perceiving model as input to training our networks. Then, we use the three-layer residual blocks to merge all the information and generate full size detection results. Besides, we build a new high quality dataset with the complex wild environment based on dataset design principles. The experimental results on VWM dataset demonstrate our algorithm improves the quality of wild animal detection and is more robustness. To encourage future works, we make the dataset and related materials open. All of these can be found on our project website[1].

5. Acknowledge

We would like to thank anonymous reviewers for their helpful comments on the paper. This work was supported by the grant of National Natural Science Foundation of China (No.U1611461), the grant of Science and Technology Planning Project of Guangdong Province, China (No.2014B090910001) and the grant of Shenzhen Peacock Plan (No.20130408-183003656).

References

[1] M. M. Cheng, N. J. Mitra, X. Huang, P. H. Torr, and S. M. Hu. Global contrast based salient region detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(3):569–582, 2015.

[2] J. H. Giraldozuluaga, A. Gomez, A. Salazar, and A. Diazpulido. Camera-trap images segmentation using multilayer robust principal component analysis. 2017.

[3] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[4] B. Hughes and T. Burghardt. Automated visual fin identification of individual great white sharks. *International Journal of Computer Vision*, 122(3):542–557, 2017.

[5] H. Li, H. Lu, Z. Lin, X. Shen, and B. Price. Inner and inter label propagation: salient object detection in the wild. *IEEE*

[1] https://chunbiaozhu.github.io/prof.cbzhu/

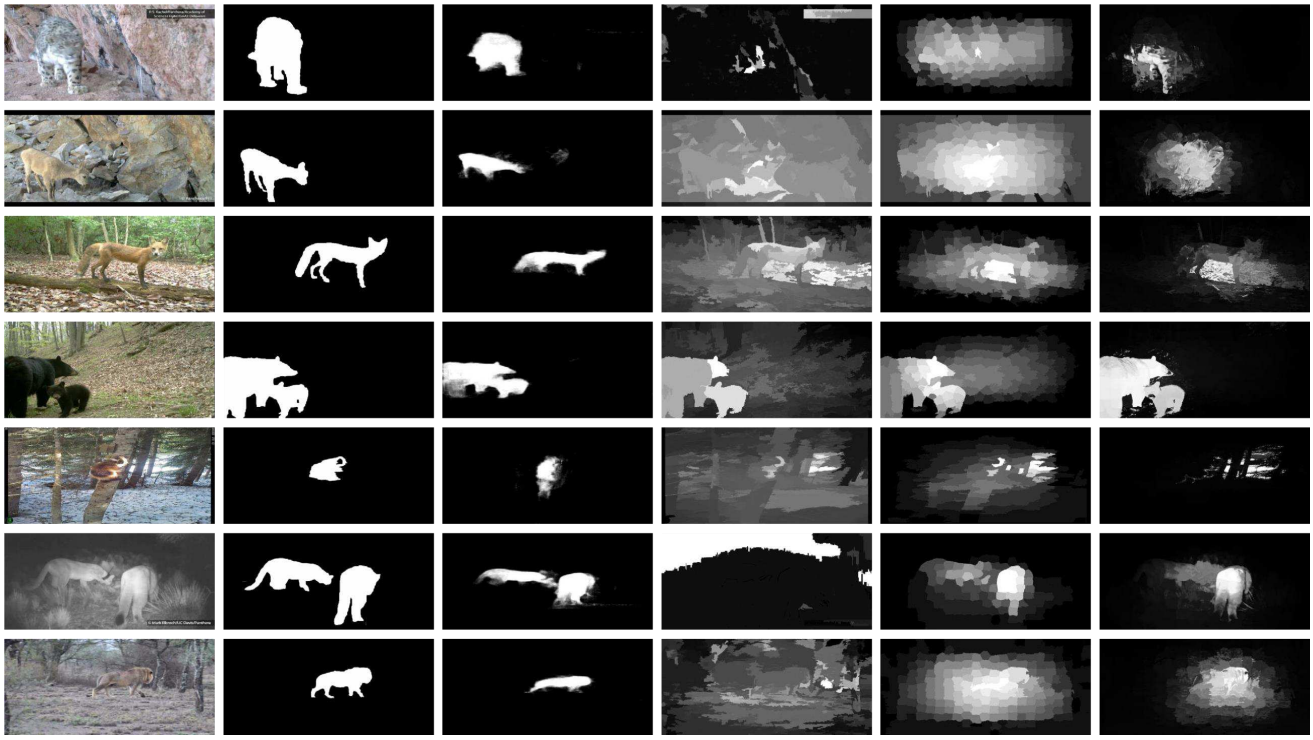

Figure 5. The visual comparison results, from left to right: Input Image, Ground Truth, TPRPN, HS, BSCA and LPS.

Transactions on Image Processing A Publication of the IEEE Signal Processing Society, 24(10):3176–3186, 2015.

[6] A. Lucchi, K. Smith, R. Achanta, G. Knott, and P. Fua. Supervoxel-based segmentation of mitochondria in em image stacks with learned shape features. *IEEE Transactions on Medical Imaging*, 31(2):474–86, 2012.

[7] Y. Qin, H. Lu, Y. Xu, and H. Wang. Saliency detection via cellular automata. In *Computer Vision and Pattern Recognition*, pages 110–119, 2015.

[8] S. Ravela, C. Yang, J. Runge, L. Gamble, K. Mcgarigal, M. Chesser, and B. Timm. Visual recapture for movement ecology at interannual timescales. 2008.

[9] J. Shi, Q. Yan, L. Xu, and J. Jia. Hierarchical image saliency detection on extended cssd. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(4):717–729, 2016.

[10] A. Shukla and S. Anand. Metric learning based automatic segmentation of patterned species. In *IEEE International Conference on Image Processing*, pages 3982–3986, 2016.

[11] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *Computer Science*, 2014.

[12] D. Sun, S. Roth, and M. J. Black. Secrets of optical flow estimation and their principles. In *Computer Vision and Pattern Recognition*, pages 2432–2439, 2010.

[13] A. Swanson, M. Kosmala, C. Lintott, R. Simpson, A. Smith, and C. Packer. Snapshot serengeti, high-frequency annotated camera trap images of 40 mammalian species in an african savanna. *Nature Scientific Data*, 2(6):150026, 2015.

[14] X. Yu, J. Wang, R. Kays, P. A. Jansen, T. Wang, and T. Huang. Automated identification of animal species in camera trap images. *Eurasip Journal on Image and Video Processing*, 2013(1):52, 2013.

[15] C. Zhu and G. Li. A three-pathway psychobiological framework of salient object detection using stereoscopic technology. In *2017 IEEE International Conference on Computer Vision Workshop (ICCVW)*, 2017.

[16] C. Zhu, G. Li, X. Guo, W. Wang, and R. Wang. *A Multilayer Backpropagation Saliency Detection Algorithm Based on Depth Mining*, pages 14–23. Springer International Publishing, Cham, 2017.

[17] C. Zhu, G. Li, N. Li, X. Guo, W. Wang, and R. Wang. An innovative saliency detection framework with an example of image montage. In *ACM MultiMedia Workshop 2017 Submission Proposal for South African Academic Participation Proceedings*, 2017.

[18] C. Zhu, G. Li, W. Wang, and R. Wang. An innovative salient object detection using center-dark channel prior. In *2017 IEEE International Conference on Computer Vision Workshop (ICCVW)*, 2017.

[19] C. Zhu, G. Li, W. Wang, and R. Wang. Salient object detection with complex scene based on cognitive neuroscience. In *Multimedia Big Data (BigMM), 2017 IEEE Third International Conference on*, pages 33–37. IEEE, 2017.