Monitoring Human-Wildlife Interactions in National Parks with Crowdsourced Data and Deep Learning

Bing Pan^{1(✉)}, Virinchi Savanapelli³, Abhishek Shukla³,
and Junjun Yin²

¹ Department of Recreation, Park, and Tourism Management, The Pennsylvania State University, University Park, PA 16802, USA
bingpan@psu.edu

² Social Science Research Institute, The Pennsylvania State University,
University Park, PA 16802, USA

³ Department of Mining Engineering, Department of Mechanical Engineering,
Indian Institute of Technology, Kharagpur, India

Abstract. This short paper summarizes the first research stage for applying deep learning techniques to capture human-wildlife interactions in national parks from crowd-sourced data. The results from objection detection, image captioning, and distance calculation are reported. We were able to categorize animal types, summarize visitor behaviors in the pictures, and calculate distances between visitors and animals with different levels of accuracy. Future development will focus on getting more training data and field experiments to collect ground truth on animal types and distances to animals. More in-depth manual coding is needed to categorize visitor behavior into acceptable and unacceptable ones.

Keywords: National parks · Human-wildlife interactions · Deep learning · Artificial intelligence · Crowdsourcing

1 Introduction

The dual mandate of the U.S. National Park Service (NPS) in “protecting natural and cultural resources” and “providing opportunities for public use and enjoyment” requires careful monitoring of visitor impact on natural and cultural resources in order to reach a delicate balance. In the summer of 2021, the visitation to U.S. national parks has reached a record level, and many popular parks face great pressure on their carrying capacity [1]. Visitor Use Management Framework has been proposed where visitor management strategies determine specific implementation and monitoring efforts. The key implementation actions are careful monitoring and survey of those stressors on ecological impacts, such as overcrowded visitors, illegal human-wildlife interactions, and altered physical environment due to visitor activities [2].

Traditional monitoring methods include ecological field assessment, visitor surveys, and participatory mapping [3], which tend to have significant staffing requirements. Fortunately, the most impacted areas are likely to be the places visitors

photographed and videotaped the most. As a result, the emergence of crowd-sourced data, e.g., numerous photos and videos voluntarily shared online by national park visitors, has provided new opportunities for cost-effectively monitoring visitor activities and resource impact. Past research related to public lands and crowd-sourced data is limited to the analysis of photo metadata to study visitors' spatial and temporal distribution, such as for estimating visitor volumes [4] and for locating most visited places within a park [5]. Computer scientists recreated location characteristics and trip itineraries from shared photos [6]. For preservation efforts, crowd-sourced YouTube videos were used to track whale and giraffe populations' movement with deep learning techniques [7]. No study to date has adopted crowd-sourced image/video data to monitor national park visitors' activities and their interactions with wildlife and natural environment. In this research, we are motivated to utilize deep learning and crowd-sourced data to monitor visitor activities and their impact on resources in national parks. This paper presents the first-stage effort of the development. In particular, we have focused on detecting and understanding human-wildlife interactions in national parks. The results can detect visitors' non-compliance activities in national parks.

2 Research Objectives and Methods

This project aims to address a main objective: *how can we monitor human-wildlife interactions from crowd-sourced data in preserved areas?* For example, we need to identify visitors and animals in pictures and determine whether the visitors are keeping a safe distance away from the animals; if not, what kind of visitor behaviors are, such as feeding, taking pictures, or touching. Deep learning technology can help detect such behavior and is applicable for assessing different types of human-wildlife interactions.

Technically, the core process of deep learning is to first train a layer of deep neural networks in order to accurately detect objects in images, e.g., humans, animals, or a scene of interest, also known as object detection [8]. After detecting those objects, additional layers of deep neural networks can be trained for specific applications in classification tasks. For instance, situational information about visitors' behaviors can be summarized into a set of predefined categories, such as feeding animals or setting up campfires. Distances can be calculated from recognized objects.

2.1 Object Detection

Object detection is a technique used to identify the locations of objects in an image; this task is considered an integral part of computer vision. Convolution neural networks (CNN) [9] are the most effective method for achieving accurate detection from image datasets. Region-based algorithms are developed to look at parts of the image with high probabilities of containing certain objects. However, because the objects vary by images, such as spatial locations and aspect ratios, a large number of potential regions should be generated, which is computationally challenging. Computer scientists avoid this problem by developing more computationally efficient algorithms like R-CNN (Regions with CNN features) and YOLO (You Only Look Once). Unlike region-based algorithms, YOLO algorithms only need a single convolutional network to predict the

bounding boxes of the objects and assign probability values of those boxes to a certain class, leading to significantly faster computation.

In comparison, the speed for R-CNN algorithms is around 5–17 fps, whereas the speed for the YOLO series is around 30–45 fps. For this reason, we chose YOLO algorithms to perform object detection. Furthermore, to ensure that we can achieve optimal results, we have applied both YOLOv4 and YOLOv4-p6 algorithms and the one with better performance is selected for further machine learning tasks. YOLOv4-p6 can work on a larger image size.

2.2 Automated Image Captioning for Classifying Visitor Behavior

After certain objects in the image or videos are detected, automated image captioning could be adopted to detect visitor behavior such as wild animal feeding. Automated image captioning is a deep learning technique used to describe an image with text. The algorithms convert pixels in an image to a sequence of words. To process both text and images, we employed convolutional neural networks (CNN) for training the model. The results are taken as inputs for an attention-based LSTM (Long Short Term Memory) network for training the model on the text.

2.3 Distance Calculation

Policies in U.S. national parks dictate the minimum distance visitors shall keep away from certain types of animals. For example, they should keep at least 100 yards away from bears and wolves and 25 yards from bison, elk, and other wildlife. For distance calculation, a simple transformation to the object detection algorithm is implemented. If both people and animals are detected in an image, bounding boxes of those objects are drawn as regions of interest (RoI). The algorithm then calculates the Euclidean distances between every two objects in an image and upscales them to the actual distance using the image's metadata. The distances are calculated between the center coordinate points of objects using OpenCV. Note that the current implementation assumes that all the objects are on the same level ground.

3 Training Setup

To enable the object detection algorithm to accurately classify the identified objects into 20 different classes of animals and humans, we obtained crowd-sourced data from online platforms such as Facebook, Instagram, Google. In total, we collected 17,515 images in YOLO format for object detection. Further, we manually annotated those images using the LabelImg software tool. The annotated data is split into two sets, with 15,918 images as the training dataset and 1,597 as the validation dataset. We have trained both the YOLOv4 and YOLOv4-p6 on the dataset.

To perform automated image captioning, we obtained 8,713 images from various online articles and national park websites. We manually annotated the images by giving captions for each image. The captions are organized into several columns. The first two columns are the image name and caption I.D.s. The next three columns are scores

representing an expert judgment (two authors of this paper) for that image-caption pair. The score range is from 1 to 4: (1) the caption does not describe the image at all; (2) the caption describes minor aspects of the image but does not describe the image; (3) the caption almost describes the image with minor mistakes; and (4) the caption describes the complete images. Finally, the dataset is split with 6,421 images for training, including 1,196 images for validation and 1,096 images for testing.

4 Results

4.1 Object Detection

Object detection performance was measured by the metric named Intersection over Union (IoU), the most popular evaluation metric used as the object detection benchmarks [9]. Whenever an object is identified in an image, a bounding box of the object is generated. The IoU is calculated as the ratio of the overlapping of the predicted bounding box with the ground truth bounding box to the union of two boxes. Its value ranges from 0 to 1, where 0 indicates no overlapping and 1 means perfect prediction.

In Table 1, A.P. represents the average precision of the model; AP50 represents the average precision when IoU is 0.5, and AP75 is when IoU is 0.75. AP50 means that the precision ratio when we consider an object is corrected identified when its IoU value is 0.50 or above. The YOLOv4-p6 algorithm has higher A.P., AP50, and AP75 values compared to the YOLOV4 algorithm. However, the average AP is around 49% and still needs further improvements.

Table 1. The results of object detection using YOLOv4 and YOLOv4-P6

Model	Image size	AP	AP50	AP75
YOLOv4	416 × 416	40.2%	57.0%	46.0%
YOLOv4-p6	1280 × 1280	49.0%	63.2%	54.0%

4.2 Image Captioning

The results for image captioning are listed in Table 2. Bilingual Evaluation Understudy Score (BLEU) is a metric to check the reference sentence with the sentence generated by the machine learning model. BLEU1, BLEU2, BLEU3, BLEU4 are the different methods based on n-gram of words. For example, BLEU1 looks at the ratio of the same single words in a testing text compared to a ground truth text, regardless of the sequence of the words; BLEU2 looks at pairs of words in two pieces of text. In general, the accuracy is around 60% with BLEU1 on the testing dataset. A picture of a moose charging at a visitor generated a captioning result: “*a group of people is capturing photos of a animal in green grass area*”. This caption can hardly describe the severity of the incidence in the picture. Thus, the algorithm still needs great improvement.

Table 2. The evaluation metrics of image captioning using BLEU scores

Dataset	BLEU1	BLEU2	BLEU3	BLEU4
Train	73.8%	58.1%	47.2%	39.1%
Validation	60.1%	38.8%	25.0%	17.4%
Test	60.6%	39.4%	25.9%	16.9%

4.3 Distance Calculation Results

We also calculated the distance from the animals in the picture to the visitors. However, since the ground truth data is hard to obtain except for the photos’ metadata, the accuracy of the calculation cannot be determined. Future studies should focus on field experiments where the distance is known, and then the machine learning model could be further tested with ground truth distance and camera metadata measurement.

5 Conclusions

This short paper summarizes the first research stage for applying deep learning techniques to capture human-wildlife interactions in national parks using crowd-sourced data. The objection detection showed promising yet unsatisfying results with an average precision of 49%. The image captioning algorithms showed a BLEU1 score of 60% on test data; the distance calculation results are inconclusive since it is hard to obtain the ground truth of the distances without field experiments. We plan to obtain more training data and conduct field experiments where the ground truth on animal type and visitor behaviors can be known. More in-depth manual coding is needed to categorize visitor behavior into acceptable and unacceptable ones.

Acknowledgments. This research is supported by a seed grant from the Institute for Computational and Data Sciences (ICDS) of the Pennsylvania State University. Virinchi Savanapelli and Abhishek Shukla implemented the machine learning models of this research.

References

1. Paoletta K (2021) Give the people what they clearly need: more National Parks. <https://www.nytimes.com/2021/08/28/opinion/national-park-nature.html>

2. Monz CA, Cole DN, Leung Y-F, Marion JL (2010) Sustaining visitor use in protected areas: future opportunities in recreation ecology research based on the USA experience. *Environ Manage* 45:551–562

3. Rice WL, Taff BD, Newman P, Zipp KY, Pan B (2020) Identifying recreational ecosystem service areas of concern in Grand Canyon National Park: a participatory mapping approach. *Appl Geogr* 125:102353

4. Sessions C, Wood SA, Rabotyagov S, Fisher DM (2016) Measuring recreational visitation at U.S. National Parks with crowd-sourced photographs. *J Environ Manage* 183:703–711

5. Barros C, Moya-Gómez B, Gutiérrez J (2020) Using geotagged photographs and GPS tracks from social networks to analyse visitor behaviour in national parks. *Curr Issues Tour* 23:1291–1310
6. Lu X, Wang C, Yang J-M, Pang Y, Zhang L (2010) Photo2trip: generating travel routes from geo-tagged photos for trip planning. In: *Proceedings of the 18th ACM international conference on multimedia*, pp 143–152
7. Berger-Wolf TY et al (2017) Wildbook: crowdsourcing, computer vision, and data science for conservation. *ArXiv Prepr. ArXiv171008880*
8. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. *ArXiv Prepr. ArXiv14091556*
9. Rezatofighi H, Tsoi N, Gwak J, Sadeghian A, Reid I, Savarese S (2019) Generalized intersection over union: a metric and a loss for bounding box regression. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp 658–666

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.