DEEP CONVOLUTIONAL NEURAL NETWORK BASED SPECIES RECOGNITION FOR WILD ANIMAL MONITORING

*Guobin Chen, Tony X. Han, Zhihai He**

{gcn38,hantx,hezhi}@missouri.edu
Univeristy of Missouri
Electitral and Comptuer Engineering Department
Columbia, MO 65203, USA

Roland Kays, and Tavis Forrester

rokays@gmail.com; ForresterT@si.edu
North Carolina State University
Department of Forestry and Environmental Resources
Raleigh, NC 27607, USA

ABSTRACT

We proposed a novel deep convolutional neural network based species recognition algorithm for wild animal classification on very challenging camera-trap imagery data. The imagery data were captured with motion triggered camera trap and were segmented automatically using the state of the art graph-cut algorithm. The moving foreground is selected as the region of interests and is fed to the proposed species recognition algorithm. For the comparison purpose, we use the traditional bag of visual words model as the baseline species recognition algorithm. It is clear that the proposed deep convolutional neural network based species recognition achieves superior performance. To our best knowledge, this is the first attempt to the fully automatic computer vision based species recognition on the real camera-trap images. We also collected and annotated a standard camera-trap dataset of 20 species common in North America, which contains $14,346$ training images and $9,530$ testing images, and is available to public for evaluation and benchmark purpose.

Index Terms— Species recognition, wild animal monitor, image classification, deep convolutional neural networks, large scale learning

1. INTRODUCTION

Our wildlife populations are increasingly imperiled as human actions are altering natural systems through aggressive resource acquisition and landscape changes. Furthermore, the urbanization of our society has decreased the personal interactions between humans and wildlife with declines in the popularity of many outdoor recreational activities, such as hunting and fishing [1]. The problematic result is that our society is causing more problems for wildlife while at the same time becoming less concerned about the well being of wildlife species and our natural systems. This creates significant hurdles in effective management of natural resources and protection of wildlife species.

The geographic scale of these conservation, ecological, and environmental issues is beyond the capability of any single study to tackle, although numerous studies have documented the phenomenon at selected sites [2]. One exception is the bird-watching networks, including Breeding Bird Survey [3], E-bird [4], and Christmas Bird Counts [5], which engage millions of citizens across the country today to report local population trends of bird species [6]. The resulting datasets on bird abundance and distribution have been the backbone of many important continental conservation programs [7, 8]. These efforts have successfully recorded the spread of invasive species [9], identified critical bird species in need of conservation actions and saved a number of species from extinction [10].

Unfortunately, most mammalian wildlife species, such as lion, deer, and tiger, are too shy to be directly observed and tracked by citizens. During the past decades, engineers and wildlife researchers have developed various technologies for professionals to monitor individual mammals, including very high frequency (VHF) radio tracking [11], satellite tracking [12], and Global Positioning System (GPS) tracking [13, 14, 15], wireless sensor networks [16, 17, 18], and animal-mounted video monitoring systems [19]. However, these efforts have been mostly carried out on a relatively small number of wildlife species by professional wildlife researchers, over a short period of time (often in the range of a few hours, days, or weeks), and over small geographical areas. Furthermore, camera and sensor data collected by different individuals and research groups are scattered in space and time, represented in various forms, and isolated from each other.

With technological advances in hardware and embedded computing, existing camera-trap technologies for wildlife monitoring (e.g. Reconyx camera systems [20]) have matured to where they are commercially available at a reasonable cost, rapidly deployable, easy to maintain, and therefore to be practically used by a large number of non-professional

*This work was partly supported by NSF under the grants DBI-1062354 and EF-1065749.

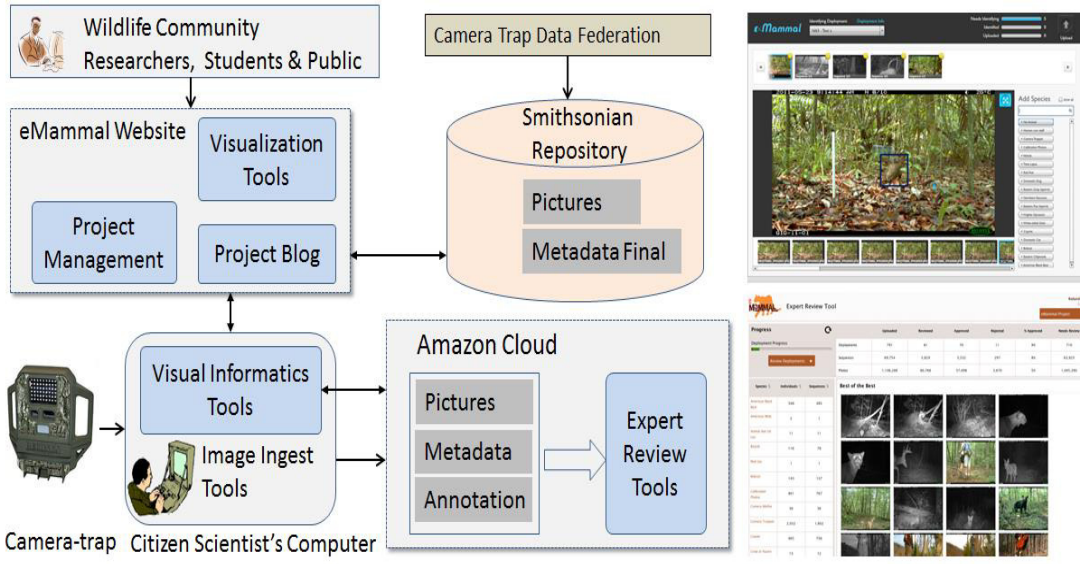

Fig. 1: The framework of eMammal cyber-infrastructure.

citizens. With the help of a group volunteers, we have been able to employed camera traps at around $1,000$ locations to capture the imagery data of wild animals so that we can analyze their behaviors and monitor the population. This on-going eMammal project has been developing a citizen scientist-based camera trap monitoring cyber-infrastructure to collect large-scale data on animal populations, species richness and diversity, and engage the public in both nature and scientific exploration.

The eMammal project leads to an biological informatics cyber-infrastructure which brings together citizen scientists and wildlife professionals to collect, analyze, and manage massive camera-trap data for collaborative wildlife research at large scales. Figure 1 shows the basic framework of eMammal cyber-infrastructure. As shown in Figure 1, the imagery data captured by camera trap need to be annotated with information including the animal location, animal category, and the moving speed. Due to the scale of the imagery data captured by the $1,000$ camera traps, manually labeling these imagery data is formidable. We therefore resort to computer vision algorithm to annotate these wild animal imagery data automatically.

With our previous work Ensemble Video Object Cut (EVOC) [21], we can segment out the object of interests, i.e. the wild animals. Therefore, to make the aforementioned eMammal platform an automatic tool for biologists, the key problem is visual species recognition. That is we need to recognize the animal category of the current image sequence captured by the camera trap so that we can analyze behavior patterns and populations of different species. We therefore proposed a novel deep convolutional neural network based species recognition algorithm for wild animal classification on these very challenging camera-trap imagery data. For the comparison purpose, we use the traditional bag of visual words model [22, 23] as the baseline species recognition algorithm. As shown in Section 4, it is clear that the proposed deep convolutional neural network based species recognition achieves superior performance. To our best knowledge, this is the first attempt to the fully automatic computer vision based species recognition on the real camera-trap images. We also collected and annotated a standard camera-trap dataset of 20 species common in North America, which contains $14,346$ training images and $9,530$ testing images, and is available to public for evaluation and benchmark purpose.

2. RELATED WORK

The deep learning algorithms [24, 25, 26] have shown their advantages in various tasks including Natural Language Processing (NLP) [27, 28], speech recognition [29] and computer vision [30, 31]. With great performances and the capability suitable for large scale learning, the deep learning algorithms provide a promising path to problems that are very difficult to traditional machine learning algorithms such as visual recognition. Recently, the deep learning based algorithm [30] achieved a dominant win over traditional algorithms including SVM, boosting algorithms, and multiple kernel learning in the Large Scale Visual Recognition Challenge 2012 (ILSVRC2012) [32]. However, there are still many issues in deep learning algorithms need to be further investigated. For example, there are several variants of deep learning algorithms and the performances of these algorithms vary drasti-

cally for different tasks. Currently there are no through investigation and summary on the comparisons of these algorithms. Besides, current Restricted Boltzman Machine (RBM) based deep learning algorithms have a very small receptive field, which works best for small input images such as 30 by 30 handwritten digit images. Considering all of the aforementioned factors, we mainly applied the deep convolutional neural network [30] with task specific variation for our species recognition purpose.

3. VISUAL SPECIES RECOGNITION ON NOISY CAMERA-TRAP DATA USING DEEP CNN

The imagery data captured by the camera traps are image sequences triggered by a motion sensor with the sequence length ranging from 6 frames to 50 frames. Our previous work, EVOC [21] is first applied to the image sequence to segment out the moving foreground. The tight bounding box around the segmented region are selected as the Region of Interest (ROI). We can then treat the species recognition problem as the image classification on the ROIs. Some sample ROI cropped out of the camera-trap images are shown in Figure 2.

Fig. 2: The sample ROIs cropped out of the camera-trap images using EVOC [21].

Since the EVOC based foreground segmentation algorithm cannot generate perfect aligned bounding box, the image classification algorithm has to tolerate the imprecision of the ROIs such as part clipping or very loose ROI. We therefore resort to two image classification algorithms: (1) Bag of visual Words (BOW) model based image classification algorithms [22, 23]; (2) Deep Convolutional Neural Network (DCNN) based image classification algorithms. These two image classification algorithms both have their own advantages and disadvantages. The BOW image classification algorithm is simple and quite robust to deformation and part clipping, but it achieves only suboptimal results. The DCNN based image classification algorithm [30] can achieve superior performance over most of the state-of-the-art image classification algorithms [32], but requires large amount of labeled training data, even if the data augmentation technique [30] is applied.

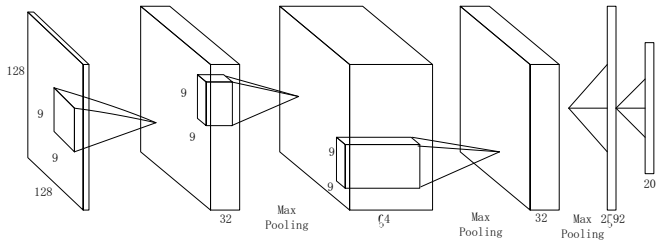

Fig. 3: The structure of the DCNN used for species recognition.

Considering the amount of training data available, we designed a DCNN with 3 convolutional layers and 3 max pooling layers. The convolutional layer has a convolutional kernel with a size of 9×9 ,while pooling layer has a kernel with a size of 2×2. The input layer size is 128×128. In the first convolutional layer,which apply 2-D convolution to the 128×128 input layer,we can get a 120×120 output matrix. Since we have 32 kernels in the first convolution layer,we can get 32 out matrices. Then we apply 2×2 max pooling. That is we use the highest value in a 2×2 block to represent that block. Then after pooling we have 32 60×60 matrices as the output of the first layer, which are the input of the second convolution layer. For each kernel in the second layer, we apply convolution to each input matrix and take average to get a output matrix. The second layer outputs 64 52×52 matrices, pooling to 64 26×26 matrices. The 3rd pooling layer outputs 32 9×9 matrices and we make it into a 2592 dimensional vector. After that is a fully connected layer and a soft max layer. The soft max layer have 20 neurons and we can use the max output among these 20 neurons to determine the label of input image. The data augmentation step [30] is also used during our training stage.

Table 1: Species Recognition Performance Comparison on Camera-trap Data.

Method	Agouti	Peccary	Paca	R-Brocket Deer	W-nosed Coati	Spiny Rat	Ocelot	R-Squirrel	Opossum	Bird spec
BOW	0.041	0.108	0.298	0.01	0.333	0.146	0.398	0.028	0.296	0.011
DCNN	0.13	0.122	0.187	0.02	0.243	0.05	0.224	0.038	0.147	0.001
Method	Tinamou	W-Tail Deer	Mouflon	R-Deer	Roe Deer	Wild Boar	R-Fox	Euro Hare	Wood Mouse	Coiban Agouti
BOW	0.397	0.69	0.647	0.746	0.038	0.246	0.001	0.143	0.746	0.055
DCNN	0.298	0.5	0.71	0.82	0.046	0.171	0.001	0.02	0.873	0.045

4. EXPERIMENTAL RESULTS

4.1. Camera-trap Dataset for Species Recognition Benchmarking

We collected and annotated a standard camera-trap dataset of 20 species common in North America, which contains $14,346$ training images and $9,530$ testing images, and is available to public for evaluation and benchmark purpose. The 20 species are: Agouti, Collared Peccary, Paca, Red Brocket Deer, White-nosed Coati, Spiny Rat, Ocelot, Red Squirrel, Common Opossum, Bird spec, Great Tinamou, White Tailed Deer, Mouflon, Red Deer, Roe Deer, Wild Boar, Red Fox, European Hare, Wood Mouse, and Coiban Agouti. The training and testing images are randomly sampled from the total collection of images including color images, gray images, and infrared images with resolutions raging from 320 by 240 to 1024 by 768. Each image icontains only one type of animal out of the aforementioned 20 categories. The accuracy on the testing images are used to benchmark different algorithms.

4.2. Species Recognition Baseline using BOW based Image Classification

We follow the famous bag-of-words model to do the classification. First, we divide the whole image into overlapping small blocks,i.e.8 by 8 block. The blocks are the ?words?. We can extract features to represent the block. Putting features from all images together, we can get a ?vocabulary? of visual words. Then according to the vocabulary, each image have a histogram of occurrence counts of words. We use the histogram to represent the image, and use linear SVM as the classifier. Since this representation ignore the spatial relation between image blocks, it is tolerance to large deformation. We use 8 by 8 blocks as visual words get a block every 3 pixels. For each block ,a 128 dimensional SIFT feature was extracted to represent the patch. We random sampled 1000000 features from all the training image and train a code book using k-means clustering. Here is the result of the BoW model with different code size: for $K = 1000, 2000, 3000$, the accuracies of the BoW model are 33.192%, 33.507%, and 33.485% respectively.

4.3. Species Recognition Results of DCNN and Bow for 20 species

With on the collected camera-trap dataset, we compared the BOW model with our DCNN algorithm for species recognition. The performance comparison is shown in Table 1. The overall species recognition accuracy of the BOW is 33.507% and the overall species recognition accuracy of the DCNN is 38.315%. We also want to emphasize that the learning capacity of the DCNN is very high and therefore the performance of the DCNN can be further improved if more training data are available. From this comparison we can find that the proposed DCNN outperforms the traditional BOW model. Although the current performance on this very challenging dataset (as shown in Figure 2) cannot meet the fully automatic requirements, we can still use the DCNN algorithm to select ambiguous data for annotation, which can alleviate the burden of the experts to large extent.

5. CONCLUSION AND DISCUSSION

We proposed a novel DCNN based species recognition algorithm. On the very challenging real camera-trap imagery data set, our DCNN based species recognition algorithm outperforms the traditional BOW based species recognition algorithm and show promising results. Although the current performance has not meet the requirements of full automation. But the most confident recognition results can already alleviate the burden of the annotation experts to a large extent. With more camera-trap data collected, we expect that the DCNN based species recognition algorithm can improve quickly due to its large learning capacity and finally achieve the goal of automatic species recognition for camera-trap data.

6. REFERENCES

[1] Shawn J. Riley, Daniel J. Decker, Jody W. Enck, Paul D. Curtis, T. Bruce Lauber, Tommy L. Brown, and Rileey S J, "Deer populations up, hunter populations down: Implications of interdependence of deer and hunter population dynamics on management," *Ecoscience*, vol. 10, pp. 455–461, 2003.

[2] L. Markovchick-Nicholls, H.M. Regan, D.H. Deutschman, A. Widyanata, B. Martin, L. Noreke, , and T.A Hunt, "Relationships between human disturbance and wildlife land use

in urban habitat fragments," *Conservation Biology*, vol. 22, pp. 99–109, January 2008.

[3] J. R. Sauer, J. E. Hines, and J. Fallon, "The north american breeding bird survey, results and analysis 1966 - 2007," USGS Patuxent Wildlife Research Center, 2008, vol. Version 5.15.

[4] eBird, http://ebird.org/content/ebird/.

[5] Christmas Bird Count, http://www.audubon.org/bird/cbc/.

[6] Stephen N. Freeman, David G. Noble, Stuart E. Newson, and Stephen R Baillie, "Modelling population changes using data from different surveys: the common birds census and the breeding bird survey," *Bird Study*, vol. 54, pp. 61–72, March 2007.

[7] William A. Link, John R. Sauer, and Daniel K. Niven, "Combining breeding bird survey and christmas bird count data to evaluate seasonal components of population change in northern bobwhite," *Journal of Wildlife Management*, vol. 72, pp. 44–51, January 2008.

[8] A. M. Pidgeon, V. C. Radeloff, C. H. Flather, C. A. Lepczyk, M. K. Clayton, T. J. Hawbaker, and R. B. Hammer, "Associations of forest bird species richness with housing and landscape patterns across the usa," *Ecological Applications*, vol. 17, pp. 1989–2010, October 2007.

[9] C. K Wikle, "Hierarchical bayesian models for predicting the spread of ecological processes," *Ecology*, vol. 84, pp. 1382–1394, 2003.

[10] J. A Veech, "A comparison of landscapes occupied by increasing and decreasing populations of grassland birds," *Conservation Biology*, vol. 20, pp. 1422–1432, 2006.

[11] L. D. Mech, *A Handbook Of Animal Radio-tracking*, Univ. of Minn. Press, Oxford, 1983.

[12] M. Williams, A. Lunsford, D. Ellis, J. Robinson, P. Coronado, and W. Campbell, "Satellite tracking of threatened species," *Argos Newsletter*, vol. 53, 1998.

[13] P. Juang, H. Oki, Y.Wang, M. Martonosi, L.-S. Peh, and D. Rubenstein, "Energy-efficient computing for wildlife tracking: design tradeoffs and early experiences with zebranet," in *Tenth International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS-X)*, San Jose, CA, 2002.

[14] Jon Young and Tiffany Morgan, *Animal Tracking Basics*, Stackpole Books, 2007.

[15] Ian A.R. Hulbert and John French, "The accuracy of gps for wildlife telemetry and habitat mapping," *Journal of Applied Ecology*, vol. 38, pp. 869–878, August 2001.

[16] I.F. Akyildiz, Y. Sankarasubramaniam W. Su, and E. Cayirci, "Wireless sensor networks: a survey," *Computer Networks*, vol. 38, pp. 393–422, March 2002.

[17] H. Gharavi and K. Ban, "Vision-based ad-hoc sensor networks for tactical operations," in *World Wireless Congress, 3G Wireless 2002*, San Francisco, 2002.

[18] Robert Szewczyk, Alan Mainwaring, Joseph Polastre, and David Culler, "An analysis of a large scale habitat monitoring application," in *In Proceedings of the Second ACM Conference on Embedded Networked Sensor Systems (SenSys)*, 2004.

[19] R. J. Moll, J. Millspaugh, J. Beringer, J. Sartwell, and Zhihai He, "Animal-borne video systems: a new era of behavioral ecology," *Trends in Ecology and Evolution*, vol. 22, pp. 660–668, November 2007.

[20] Reconyx PC85 RapidFire Professional Color IR camera system, http://www.reconyx.com/shop/Professional_Research_Camera_Traps/56.

[21] Xiaobo Ren, Tony X. Han, and Zhihai He, "Ensemble video object cut in highly dynamic scenes," *2013 IEEE Conference on Computer Vision and Pattern Recognition*, vol. 0, pp. 1947–1954, 2013.

[22] David M. Blei, Andrew Y. Ng, and Michael I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003.

[23] Fei-Fei Li and Pietro Perona, "A bayesian hierarchical model for learning natural scene categories," in *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 2 - Volume 02*, Washington, DC, USA, 2005, CVPR '05, pp. 524–531, IEEE Computer Society.

[24] Geoffrey E. Hinton, "Deterministic boltzmann learning performs steepest descent in weight-space," *Neural Comput.*, vol. 1, no. 1, pp. 143–150, Mar. 1989.

[25] Yann Lecun, Lon Bottou, Yoshua Bengio, and Patrick Haffner, "Gradient-based learning applied to document recognition," in *Proceedings of the IEEE*, 1998, pp. 2278–2324.

[26] Clement Farabet, Camille Couprie, Laurent Najman, and Yann LeCun, "Learning hierarchical features for scene labeling," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1915–1929, 2013.

[27] Antoine Bordes, Xavier Glorot, Jason Weston, and Yoshua Bengio, "Joint learning of words and meaning representations for open-text semantic parsing," in *AISTATS*, 2012, pp. 127–135.

[28] Richard Socher, Eric H. Huang, Jeffrey Pennington, Andrew Y. Ng, and Christopher D. Manning, "Dynamic pooling and unfolding recursive autoencoders for paraphrase detection," in *NIPS*, 2011, pp. 801–809.

[29] Dong Yu, Li Deng, and Frank Seide, "The deep tensor neural network with applications to large vocabulary speech recognition," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 21, no. 2, pp. 388–396, 2013.

[30] Alex Krizhevsky, Ilya Sutskever, and Geoff Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, P. Bartlett, F.C.N. Pereira, C.J.C. Burges, L. Bottou, and K.Q. Weinberger, Eds., pp. 1106–1114. 2012.

[31] Koray Kavukcuoglu, Pierre Sermanet, Y-Lan Boureau, Karol Gregor, Michal Mathieu, and Yann LeCun, "Learning convolutional feature hierarchies for visual recognition.," in *NIPS*, 2010, pp. 1090–1098.

[32] Large Scale Visual Recognition Challenge 2012 (ILSVRC2012), http://www.image-net.org/challenges/LSVRC/2012/.