Wildlife Action Recognition using Deep Learning

Weining Li, Sirnam Swetha, and Dr. Mubarak Shah
University of Central Florida
Orlando, Florida 32816

***Abstract*—Action recognition is the task of identifying the action(s) performed in a video. Till date most of the action recognition systems are human-centric. In this work, we propose to perform action recognition in wildlife which is not much explored. This involves many difficulties that come with recognizing different movements between different species of animals, unlike human actions which typically look similar across individuals. In this work, we introduce an animal action dataset which can be used for generic animal action recognition. We evaluate the proposed dataset on multiple approaches - the I3D model, a fusion network of I3D and VGG for scene semantic features, and a hierarchy of networks. We evaluate the performance of the baseline approaches and report accuracy.**

I. Introduction

Try imagining a cheetah and a crocodile running, do you find any resemblances? Unlike action recognition for humans where the action "running" looks similar for most individuals, animals are specialized based on their class, order, suborder, and family of the species, making it challenging to learn actions in wildlife. In this work, we aim to develop a system to learn animal action recognition in the wild without any manual labelling effort. A use case might be for ecologists who are interested in species interactions, behavioral understanding and how one species influence the occurrence of another. It is tedious and difficult for human to process all the data from camera traps. A system like this can help extract information such as species presence and their behaviors which can significantly reduce human efforts.

There have been an increase in interest for animal action recognition systems along with development in other related deep learning areas such as animal re-identification [1, 3], animal behavior understanding for specific species [5], and animal recognition [4] and localization [2]. Recently, researchers have conducted competitions like iWildCam 2020 [11] to solve problems in this space. This is also important as it has large scale practical applications, for example the documentaries released by National geography and BBC Earth intended to study and analyze animals and their behavioural dynamics require a lot of manual pre-processing which can be aided with computer vision systems.

We propose a dataset consisting of different species of animals, including terrestrial, aquatic, and airborne, performing both similar and unique actions. The videos typically include various environments, camera motion, and other factors that make the dataset challenging overall. We also utilize this dataset to explore three different approaches for animal action recognition - a normal I3D model, a fusion network that combines both generic features and scene semantic features, and a hierarchy of networks that first groups the dataset by action and then separate by animals and compare their performance.

Fig. 1. Example frames taken from 4 different categories of the animal action dataset

II. Related Work

[6] introduces the I3D model, an architecture for video action classification. I3D was adopted by expanding the filters and pooling kernels of a 2D architecture into 3D (NxN is expanded to NxNxN) in order to process the spacial-temporal information in videos. The 3D filters were bootstrapped from 2D filters on already trained models and ensured that the convolutional filter responses stay the same. The paper also contributes the idea of transfer learning from Kinetics dataset to other video tasks. It was shown that, after pre-training, I3D models were able to reach 80.9% on HMDB-51 [10] and 98.0% on UCF-101 [9].

In recent years, there have been works relating to animals and working with camera trap data. However, most of them are limited. Many of them, like [1], [2], [3], and [5], deals with extracting information from single images instead of videos. Most existing explorations at animal action recognition are focused on different actions performed by the same animal species like explained in [4]. In terms of datasets, all of the existing ones also correspond to only a single species as shown

Animal	Actions
elephant	rest, eat, walk, swim
wolf	rest, eat, walk, swim
lion	rest, eat, walk, swim
chicken	rest, eat, walk, swim
alligator	rest, eat, walk, swim
rabbit	rest, eat, walk, hop, swim
deer	rest, eat, walk
giraffe	rest, eat, walk
ostrich	rest, eat, walk
kangaroo	rest, eat, walk, hop
snake	rest, eat, slithering
bear	rest, eat, walk
horse	rest, eat, walk
penguin	rest, eat, walk
shark	rest, eat, swim
frog	rest, eat, swim, hop
octopus	rest, eat, swim
turtle	rest, eat, swim
fish	rest, eat, swim
crab	rest, eat, walk
lobster	rest, eat, walk
ray	rest, eat, swim
flying fish	rest, eat, swim
eagle	rest, eat, fly
seagull	rest, eat, fly
bat	rest, eat, fly
pelican	rest, eat, fly
butterfly	rest, eat, fly
dragonfly	rest, eat, fly
hummingbird	rest, eat, fly
bee	rest, eat, fly
flying squirrel	rest, eat, fly, walk

TABLE I
ANIMAL AND THEIR CORRESPONDING ACTIONS USED TO FORM THE CLASSES OF THE DATASET

in [3]. Therefore, we propose a dataset and system that pertains to learning about a diverse set of species in videos.

III. DATASET

The dataset is drafted to have a total of 106 classes. As shown in Table I, it is constructed from thirty-two (32) animal categories with three to four action categories for each animal, containing 100 videos per class. For evaluation, we chose a subset of the dataset with 18 categories, including alligator eat, alligator rest, chicken eat, chicken rest, chicken swim, chicken walk, elephant eat, elephant rest, elephant swim, elephant walk, lion eat, lion rest, lion swim, lion walk, wolf eat, wolf rest, wolf swim, wolf walk.

The videos are downloaded from YouTube [12]. We collected video URLs through YouTube's Data API. With those URLs, we downloaded the videos using the python package pytube. As shown in Figure 1, there are various backgrounds, camera motion, different lighting conditions, bad recording methods, etc. to make the dataset more challenging.

IV. EXPERIMENTS

In this section we discuss the three approaches that are used for animal action recognition and report their results.

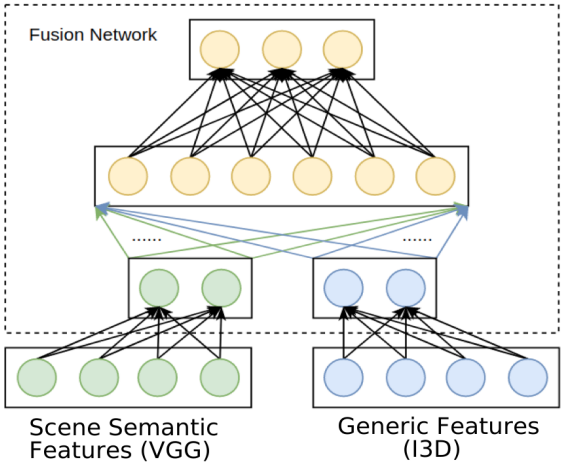

Fig. 2. Generic and scene semantic feature fusion network

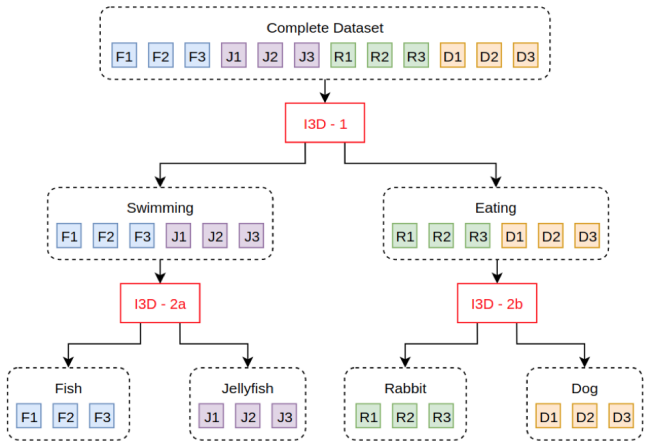

Fig. 3. Hierarchy of networks where I3D-1 handles grouping by action and I3D-2a/b handles separation by animals. Here we show two sample actions to depict the hierarchy

A. I3D

As a first baseline we propose to use I3D model. I3D is a common feature extracting model for video processing and is used for video and action classification tasks. The dataset is passed to the I3D model directly for training and evaluation.

B. Fusion Network

We also explore the role of scene semantics for action recognition as works like [8] and [9] have shown them to help the performance. I3D is used to extract generic features and VGG is used to extract scene semantic features from each clip. The two streams are then combined in the fusion network, which consists of two hidden layers and one output layer. This approach, as detailed in [8], can achieve improvements in supervised activity and video categorization.

C. Hierarchy of Networks

Intuitively, dividing the system into layers that each handle its own specific role may help improve the overall performance

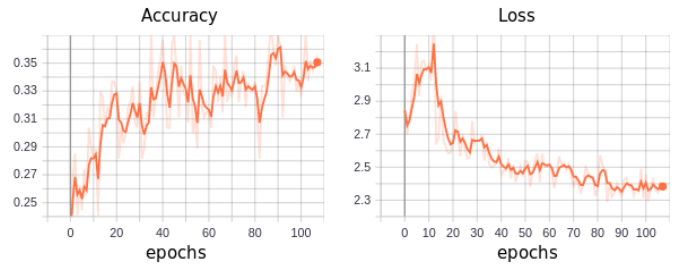

Fig. 4. Accuracy and loss graph for the I3D-only approach over the number of epochs

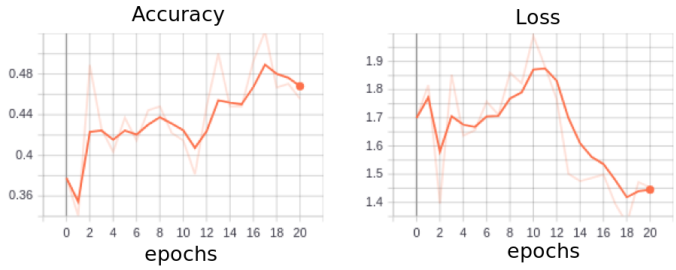

Fig. 5. Accuracy and loss graph for the first layer of the hierarchy approach over the number of epochs

of action recognition. We explore a hierarchy containing two layers of networks. As shown in Figure 3, the first layer consists of an I3D model that groups the dataset by action. The second layer consists of several I3D models that each handle a specific action. Each I3D model in the second layer will take in the videos grouped under the action and further separate the videos based on animals.

D. Results

Figure 4 shows that the I3D-only approach had about 36% accuracy. The hierarchy of networks had about 48% accuracy in the first layer and about 33% accuracy overall as shown in Figure 5 and 6 . The fusion network had about 22% accuracy, but it was ran without using pre-trained weight initialization, hence it is not directly comparable to the other approaches.

V. Conclusion

We introduced a dataset of which includes 106 categories of animal actions. We ran animal action recognition experiments on a subset of the dataset using three approaches: an I3D-only model, a fusion network using I3D and VGG without pre-trained weights, and a hierarchy of I3D models. We found that the I3D-only model had the best performance with 36% accuracy, which is still low as anticipated due to the challenges. Some further exploration could be: (1) loading pre-trained weights into the models for the fusion network and (2) implementing the hierarchy of networks in an end-to-end fashion to see if that will improve their performance.

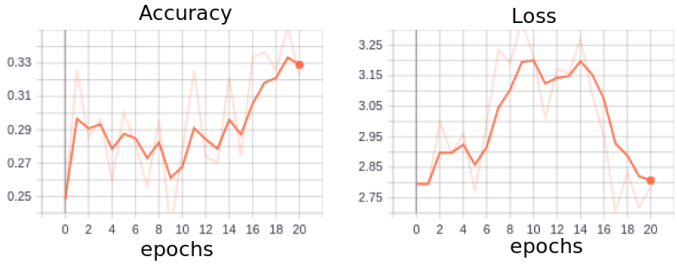

Fig. 6. Accuracy and loss graph for the best performing I3D model in the second layer of the hierarchy approach over the number of epochs

Approach	**Accuracy**
I3D	36%
Hierarchy of Networks	33%

TABLE II
ANIMAL ACTION RECOGNITION ACCURACY ON THE PROPOSED BASELINES

References

[1] M. S. Norouzzadeh, A. Nguyen, M. Kosmala, A. Swanson, M. S. Palmer, C. Packer, and J. Clune, "Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning," in *Proceedings of the National Academy of Sciences*, 2018.

[2] S. Schneider, G. W. Taylor, and S. C. Kremer, "Deep learning object detection methods for ecological camera trap data," in *2018 15th Conference on Computer and Robot Vision (CRV)*, 2018.

[3] S. Schneider, G. W. Taylor, S. Linquist, and S. C. Kremer, "Similarity learning networks foranimal individual re-identification-beyond the capabilities of a human observer," in *Proceedings of the IEEE Winter Conference on Applications of Computer Vision Workshops (WACVW)*, 2020.

[4] G. George, A. Namdev, and S. Sarma, "Animal action recognition: Aanalysis of various approaches," *Internation Journal of Engineering Sciences Research Technology*, 2018.

[5] C.-A. Brust, T. Burghardt, M. Groenenberg, C. Kading, H. S. Kuhl, M. L. Manguette, and J. Denzler, "Towards automated visual monitoring of individual gorillas in the wild." in *Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCVW)*, 2017.

[6] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset." in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[7] Z. Wu, Y. Fu, Y.-G. Jiang, and L. Sigal, "Harnessing object and scene semantics for large-scale video understanding," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[8] C. Xu, S.-H. Hsieh, C. Xiong, and J. J. Corso, "Can humans fly? action understanding with multiple classes of actors," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[9] K. Soomro, A. R. Zamir, and M. Shah, "Ucf101: A dataset of 101 human action classes from videos in the wild," 2012.

[10] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "Hmdb: A large video database for human motion recognition," 2012.

[11] S. Beery, E. Cole, A. Gjoka, D. Morris, and S. Yang, "iwildcam 2020," https://sites.google.com/view/fgvc7/competitions/iwildcam2020.

[12] "Youtube," http://www.youtube.com.