Automated Animal Identification and Detection of Species

Utkarsh Rastogi
AIT-CSE
Chandigarh University
Punjab, India
utkarshrastogi14@gmail.com

Sarthak Singh
AIT-CSE
Chandigarh University
Punjab, India
sarthaksinghsadana@gmail.com

Dr. Vijay Bhardwaj
AIT-CSE
Chandigarh University
Punjab, India

***Abstract*—The monitoring of biodiversity is using image sensors more and more; each research produces dozens or even millions of images. One major obstacle to the development of this sector is effectively recognising the species that each photograph captures. Here, we provide an automated technique for identifying species in animal photos taken with remote camera traps. We begin our approach by cropping the backdrop from the photos. Then, we employ enhanced sparse coding spatial pyramid matching (ScSPM), which creates global features using weighted sparse coding and max pooling with a multi-scale pyramid kernel. The images are then classified using a linear support vector machine algorithm. Dense SIFT descriptor and cell-structured LBP (cLBP) are extracted as the local features. For animal localization, the majority of current systems mostly rely on human input, which does not scale well to big datasets. Specifically, we leverage the recently announced Faster-RCNN object identification framework to effectively recognize animals in photos, automating the detection process while maintaining resilience to blur, partial occlusion, lighting, and position fluctuations. In order to identify the individuals, we additionally extract features from the animal's flank using AlexNet and train a logistic regression (or linear SVM) classifier. Our system is mainly tested and assessed on a dataset of camera trap tiger photos, which consists of photographs with different illumination, animal attitude, size, and overall image quality.**

***Index Terms*—Animal Identification and Species Detection, Artificial Intelligence, Convolution Neural Networks, Data Collection and Preprocessing, Data Augmentation, Deep Learning Frameworks, and Machine Learning Algorithms.**

I. Introduction

Keeping an eye on biodiversity is crucial for our civilization, particularly regarding how land use and climate change affect wild populations. In particular, visual sensors that capture images of animals that move across their field of view—such as camera traps—offer a viable method for gathering the spatiotemporal data [1] at scales required to meet this issue. It is still difficult to interpret the massive amounts of photos produced by these investigations to identify the species of animals that were observed.

The development of computer vision methods such as posture estimation, facial expression recognition, and object identification and localization has made it possible for ecologists and researchers to automate wildlife monitoring through the systematic use of visual pattern-matching algorithms. Currently, every camera-based study of wildlife [2] [3] uses a manual method in which scientists look at each picture to determine which species are captured. This is an enormous effort for research gathering tens or even hundreds of thousands of photos.

Deep learning has recently become a potent tool for handling a variety of identification problems, including natural language processing, facial recognition, and human speech recognition. While deep learning has long been used to analyze audio signals, the majority of earlier research has been on context-gathering through human speech analysis. There haven't been many attempts to categorize various animal species using deep learning in AAIDS. Our goal is to reduce this gap by using deep learning techniques [4] to create and execute an auditory classification framework for AAIDS. A common deep learning technique used in high-level representative feature learning is the convolutional neural network (CNN). To be more precise, the input data allows CNN to extract the local spatial coherence.

Through the creation and use of automated animal identification and species detection technologies [5], the area of ecology and wildlife conservation has undergone a profound transformation in an era of quick technical innovation. This ground-breaking strategy transforms how we keep an eye on, research, and safeguard Earth's many and sometimes elusive animals by utilizing cutting-edge technology like artificial intelligence [6] [7], computer vision, and sensor networks. For academics, wildlife enthusiasts, and conservationists alike, automated animal identification and species recognition [8] represent a crucial junction of science, technology, and conservation.

Historically, it has been difficult, time-consuming, and frequently dependent on manual observation and data collecting to identify and monitor animal species in their natural environments. This approach has inherent drawbacks since the researched species may be disturbed and it is frequently unfeasible in distant or rugged terrain. Large-scale monitoring initiatives are expensive and resource-intensive since they also require a lot of human resources.

Systematic methods for detecting and identifying species of animals are now available. These remedies use a variety of technologies to overcome these constraints. Computer vision, a subfield of artificial intelligence that enables machines to

perceive and process visual data much like the human visual system, is one of this field's main foundations. These systems analyze photos and videos taken by cameras, drones, or other sensing equipment, analyze the data using cutting-edge algorithms and machine learning models, and then autonomously identify and categorize animals [9], accurately differentiating between different species.

This technological advancement has significant repercussions. With today's improved efficiency and cost-effectiveness, conservationists may perform wildlife surveys that cover more regions in less time and with less negative environmental impact. Comprehensive data on animal populations, behaviors, and habitats may be gathered by researchers to provide a fuller knowledge of ecological dynamics and the effects of environmental changes. Additionally, these systems offer a priceless tool for the early identification of threatened or endangered species, enabling prompt action and conservation.

We will dig into the technology that supports this subject as we explore automated animal identification and species recognition, from the use of machine learning techniques to the deployment of sensor networks and cutting-edge camera traps. We will also look at real-world applications in several areas, such as habitat protection, animal research, anti-poaching initiatives, and biodiversity monitoring. We will also take into account the privacy and ethical issues that these technologies raise, highlighting the necessity for their implementation to be responsible and long-lasting.

We will also explore the fascinating world of automated animal identification and species recognition, learning how this synthesis of biology and technology promises to transform how we view animals and provide fresh hope for the preservation of the planet's priceless and unique ecosystems.

II. Related Work

One significant issue that has not yet been solved is the identification of species using remote camera photos. There are several techniques for recognizing general objects in the field of computer vision. Combining spatial pyramid matching [14] (SPM) with max pooling allows for the translation invariance of an animal's body in addition to modeling the spatial arrangement of local image elements. The SPM kernel proves to be quite efficient in practice despite being straightforward to create. Modeling local features and creating an overcomplete dictionary that can sparsely represent the local features have both been accomplished with the help of sparse coding. Hard assignment and vector quantization may not produce the same outcomes as sparse coding.

CNNs have been used to automatically learn discriminating characteristics from the data in order to localize chimpanzee faces. They are also known to be resilient against occlusion and position fluctuations. Using the Snapshot Serengeti dataset, which consists of 3.2 million camera trap pictures, a variety of CNN architectures, including Alexnet, VGGnet, and ResNet, were employed to categorize 48 mammal species with a 96

In species-distribution modeling, decision trees, widely-used linear models, fluctuating logic, evolutionary algorithms for rule generation, maximum entropy techniques [13], and random forests are the typical machine-learning tools. The results of carefully selected and extracted photos containing the whole animal form are consistently quite accurate. The findings of the traditional machine learning techniques, including Support Vector Machine (SVM), sparse weighted dictionary education coding, local-structured binary patterns, and SIFT combination, were reported. However, if the animal is already known to you, this traditional algorithm needs certain unique properties. However, before being fed to the traditional machine learning method, the animal photos are also manually cropped to choose only the entire animal shape. They are mostly used for classification and individual identification.

We will be training a model to discuss and find a solution for the same as easily as we can. We will be using various types of convolutional neural networks to determine the outcome of the model just by importing the data into it and providing you the its identity.

III. Proposed Methodology

Depending on the requirements of the job and the data at hand, researchers can create unique CNN architectures or use pre-existing ones for developing Convolutional Neural Networks (CNNs) for automated animal identification and species recognition. We will be showing the architectural diagram for the layers of the convolutional neural network [10]. The following are some popular CNN architectures for image classification that may be modified to identify different kinds of animals:

VGGNET: Visual Geometry Group [11], or VGG for short, is a multi-layered, conventional deep Convolutional Neural Network (CNN) architecture. With VGG-16 or VGG-19, which are composed of 16 and 19 convolutional layers, the term "deep" refers to the quantity of layers. Innovative object identification models are built on top of the VGG architecture. The VGGNet, designed as a deep neural network, outperforms baselines on a wide range of tasks and datasets, going beyond ImageNet. Furthermore, it remains one of the most widely used image recognition systems to this day.

The VGG16 model in ImageNet attains around 92.7% top-5 test accuracy. Over 14 million images are in the ImageNet collection, which is divided into over 1000 categories. Additionally, it was one of the most popular models that was submitted to ILSVRC-2014. By replacing the enormous kernel-sized filters with several 3x3 kernel-sized filters, it performs noticeably better than AlexNet. Over several weeks, Nvidia Titan Black GPUs were utilised to train the VGG16 model.

RESNET: Rather than learning unreferenced functions, Residual Networks, or ResNets [12], learn residual functions with reference to the layer inputs. Relative nets allow these stacked layers to match a residual mapping rather than assuming that each one of them directly fits a specified underlying mapping. To create a network, they build residual blocks on

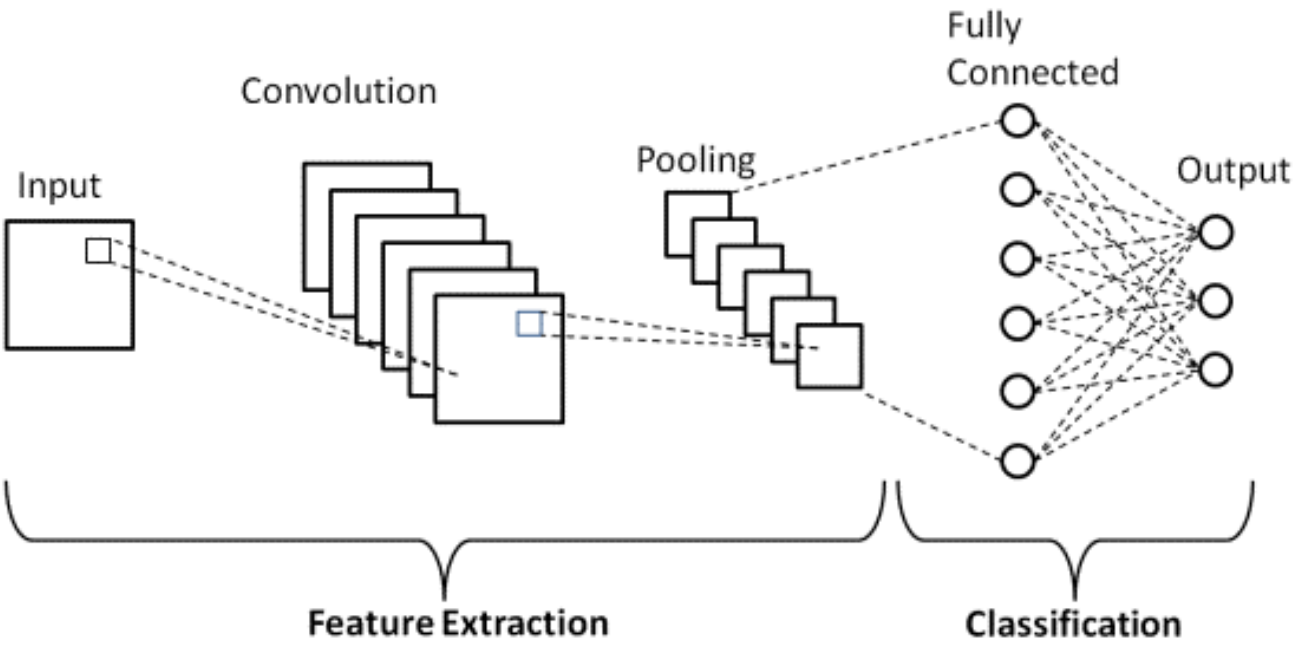


Fig. 1: CNN flow diagram

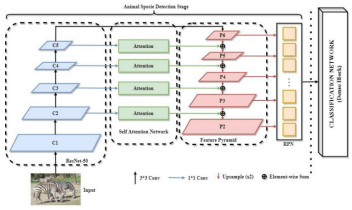


Fig. 2: Animal Species Detection Stage Flowchart

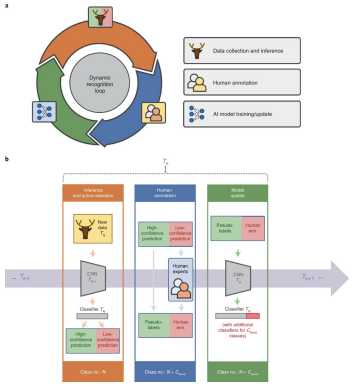


Fig. 3: Data Collection Flowchart

top of one another; for example, a ResNet-50 uses fifty layers by stacking these blocks.

One of the well-known vanishing gradient problems is one that ResNets resolves. This is due to the fact that gradients from the point at which the loss function is calculated readily drop to zero after several chain rule applications when the network is too deep. As a result, there is no learning taking place since the weights never update their values. Gradients from later layers to initial filters can flow straight through the skip connections when using ResNets.

GOOGLENET: A convolutional neural network with 22 layers is called GoogLeNet [15]. A trained version of the network trained on Places365 or ImageNet data sets may be loaded. With the help of ImageNet, the network was trained to classify photos into 1000 item categories, including several animals and keyboards, mice, and pencils. Similar to the network trained on ImageNet, the Places365 network classifies photos into 365 distinct location types, including field, park, runway, and lobby. For a large variety of pictures, these networks have learned several feature representations. The input picture size for both pre-trained networks is 224 by 224.

The inception module, which serves as the foundation for the design, is the primary innovation of GoogLeNet. Multiple convolutional layers with various kernel sizes are concatenated to collect features at different scales simultaneously in an inception module. Because of its parallelism, GoogLeNet is incredibly efficient in picture identification tasks, capturing both high-level characteristics and fine-grained details. Before using bigger convolutional kernels, each inception module performs dimensionality reduction using 1x1 convolutions, which lowers the computational cost. This prevents overfitting and saves computation in addition to serving as a regularizer.

MOBILENET: Google has made their MobileNet computer vision model open-source. It is intended to be used for classifier training. It creates a lightweight deep neural network by drastically reducing the number of parameters in comparison to other networks through the use of depthwise convolutions.

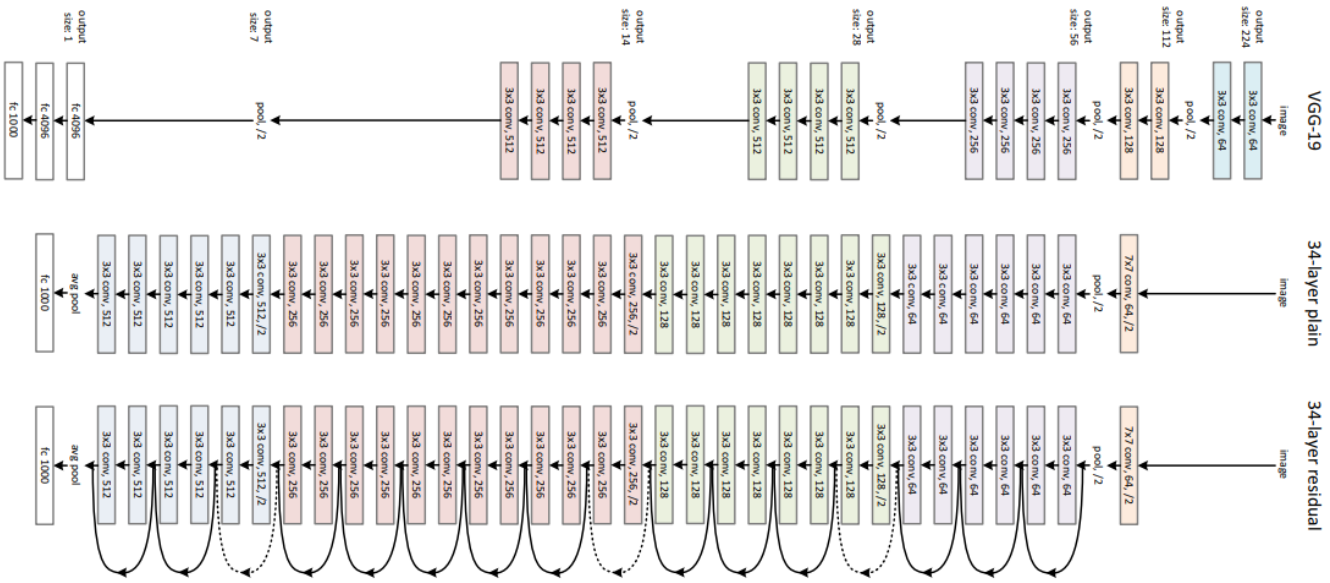


Fig. 4: Convolutional layers Architecture

Tensorflow's first mobile computer vision model is called MobileNet [16].

In comparison to other networks with conventional convolutions and the same depth in the nets, it employs depthwise separable convolutions to drastically reduce the number of parameters. Lightweight deep neural networks are the outcome of this. A family of TensorFlow computer vision models called MobileNets [17] is focused on mobile devices and is intended to optimize accuracy while taking into account the limited resources of embedded or on-device applications. MobileNets are low-power, low-latency models that are sized and configured to satisfy different use cases' resource requirements. They can serve as a foundation for segmentation, embedding, detection, and classification.

IV. Experimental Results

We used pictures of animals taken using motion-sensitive camera traps (Reconyx RC55, PC800, and HC500, Holmen, WI, USA), which, when activated by an infrared motion sensor, produce sequences of 3.1 Megapixel JPEG photos at a rate of around one frame per second. Using an infrared flash—which most animals cannot see—grayscale photographs are taken at night while color images are taken during the day. Images from temperate forests, heathlands, and tropical rainforests were utilized. Since we did not alter the data set to make identification easier, it contains many of the usual issues with camera trapping data, such as situations in which the animal is too tiny or obscured by foliage. We train a logistic regression classifier to categorize individuals using the conv3 features of AlexNet. We create five random splits for every dataset, using 25

The model for the Automated Animal Identification and Detection of Species is implemented through a code with various libraries and neural networks, and after adding the

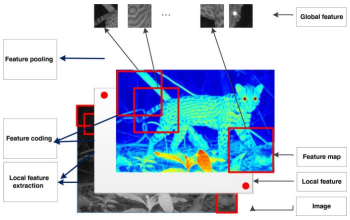


Fig. 5: Identification Result

dataset, the implementation takes place while scanning the whole document of the dataset, and a result is obtained which shows the image classification of the animal with a graph and physical attributes.

The data summary for the implementation is also shown with all the convolutional layers within the project.

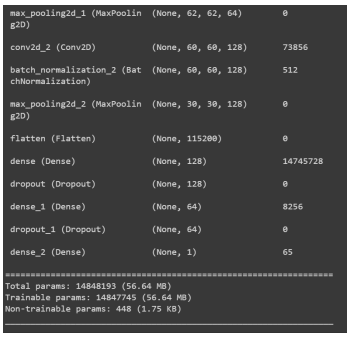
```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 conv2d (Conv2D)             (None, 254, 254, 32)      896

 batch_normalization (Batch  (None, 254, 254, 32)      128
 Normalization)

 max_pooling2d (MaxPooling2  (None, 127, 127, 32)      0
 D)

 conv2d_1 (Conv2D)           (None, 125, 125, 64)      18496

 batch_normalization_1 (Bat  (None, 125, 125, 64)      256
 chNormalization)

 max_pooling2d_1 (MaxPoolin  (None, 62, 62, 64)        0
 g2D)

 conv2d_2 (Conv2D)           (None, 60, 60, 128)       73856

 batch_normalization_2 (Bat  (None, 60, 60, 128)       512
 chNormalization)

 max_pooling2d_2 (MaxPoolin  (None, 30, 30, 128)       0
 g2D)
```

Fig. 6: Sequential Model Data

The following subsystems, which were previously shown in Figure 1, are designed in order to assess the effectiveness of the suggested network system. Now let's introduce and talk about each component individually. The most important component of the system for coexisting animals is the animal

detection subsystem. Only a few sensors are required to continually monitor animal crossing occurrences because energy is a primary concern for distant sensors. These sensors detect movement, pressure, vibration, and other environmental factors. In the meantime, it's important to distinguish legitimate animal activities from disruptions brought on by the wind and other situations that could raise false alarms, such as big birds. Prior research has demonstrated that the best sensors for this application are those that are based on fiber or infrared.

Animals are recognized by sensors that activate other sensors, such as a camera, to take pictures. These photos must then be analyzed to determine the size, position, speed, quantity, direction, and other specific characteristics of the animals. As we previously stated, the capacity of sensor nodes and the network's ability to carry data without congestion determine whether raw data is processed at the origin (end devices) or at higher levels. In the case of photos, the processed data must be transferred to fog nodes, which can reprocess it to determine whether or not to send an alert message, even if a competent end device with multiple sensors could process the raw data.



max_pooling2d_1 (MaxPooling2D)	(None, 62, 62, 64)	0
conv2d_2 (Conv2D)	(None, 60, 60, 128)	73856
batch_normalization_2 (BatchNormalization)	(None, 60, 60, 128)	512
max_pooling2d_2 (MaxPooling2D)	(None, 30, 30, 128)	0
flatten (Flatten)	(None, 115200)	0
dense (Dense)	(None, 128)	14745728
dropout (Dropout)	(None, 128)	0
dense_1 (Dense)	(None, 64)	8256
dropout_1 (Dropout)	(None, 64)	0
dense_2 (Dense)	(None, 1)	65

Total params: 14848193 (56.64 MB)
Trainable params: 14847745 (56.64 MB)
Non-trainable params: 448 (1.75 KB)

Fig. 7: Sequential Model Data 2

The results obtained from the implementation of the accuracy and precision of the model are shown as follows:

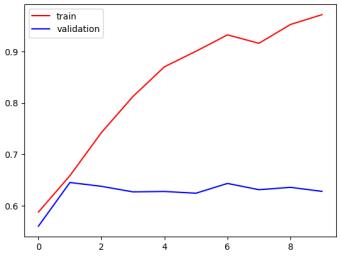


Fig. 8: Accurate Validation Graph

The camera-trapped sequences have a brief duration of around 10 frames each sequence and a low frame rate of 1 frame per second. The agouti, in which the leaves hung in the breeze, is seen in successive shots in the first two rows. The lighting drastically changes if the peccary suddenly moves in close proximity to the camera, blocking off a significant amount of light. This is not a situation that the standard motion detection algorithm can handle adequately.

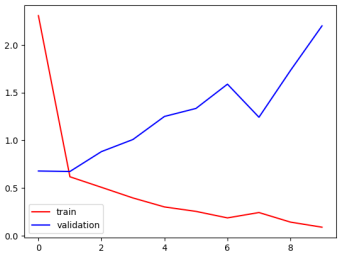


Fig. 9: Loss Validation Graph

So, in order to obtain adequate results for the dataset used we have implemented the model with the following convolutional architectural layers used above.

V. Conclusions

This work presents the design and implementation of an auditory classification system for AAIDS based on CNN. We provide a multi-view CNN architecture that includes three convolution operations with three distinct filter lengths in parallel to extract short-, middle-, and long-term information simultaneously, therefore improving accuracy in noisy environments. The suggested system performs much better than earlier approaches, according to extensive assessments of two real datasets [18].

We have demonstrated that computer vision science object identification algorithms can be applied to recognize and identify wild animals on sequences of images captured by camera traps in the field, which are known for having a high degree of noise and clutter. Even if certain species share the same ontology, the suggested approach is able to identify subtle distinctions between them.

References

[1] Shekhar, Shashi, et al. "Spatiotemporal data mining: A computational perspective." ISPRS International Journal of Geo-Information 4.4 (2015): 2306-2338.

[2] Brooks, Robert T. "Assessment of two camera-based systems for monitoring arboreal wildlife." Wildlife Society Bulletin (1973-2006) 24.2 (1996): 298-300.

[3] Miller, Anna B., Yu-Fai Leung, and Roland Kays. "Coupling visitor and wildlife monitoring in protected areas using camera traps." Journal of outdoor recreation and tourism 17 (2017): 44-53.

[4] Ghosh, Swarnendu, et al. "Understanding deep learning techniques for image segmentation." ACM computing surveys (CSUR) 52.4 (2019): 1-35.

[5] Nguyen, Hung, et al. "Animal recognition and identification with deep convolutional neural networks for automated wildlife monitoring." 2017 IEEE international conference on data science and advanced Analytics (DSAA). IEEE, 2017.

[6] Oliver, C. Ryan, et al. "A platform for artificial intelligence based identification of the extravasation potential of cancer cells into the brain metastatic niche." Lab on a Chip 19.7 (2019): 1162-1173.

[7] Congdon, Jenna V., et al. "The Future of Artificial Intelligence in Monitoring Animal Identification, Health, and Behaviour." Animals 12.13 (2022): 1711.

[8] Nguyen, Hung, et al. "Animal recognition and identification with deep convolutional neural networks for automated wildlife monitoring." 2017 IEEE international conference on data science and advanced Analytics (DSAA). IEEE, 2017.

[9] Yu, Xiaoyuan, et al. "Automated identification of animal species in camera trap images." EURASIP Journal on Image and Video Processing 2013.1 (2013): 1-10.

[10] O'Shea, Keiron, and Ryan Nash. "An introduction to convolutional neural networks." arXiv preprint arXiv:1511.08458 (2015).
[11] Sathish, R., and P. Ezhumalai. "Enhanced sentimental analysis using visual geometry group network-based deep learning approach." Soft Computing 25.16 (2021): 11235-11243.
[12] Zhou, Meilun, et al. "Improving animal monitoring using small unmanned aircraft systems (sUAS) and deep learning networks." Sensors 21.17 (2021): 5697.
[13] Gull, Stephen F., and John Skilling. "Maximum entropy method in image processing." Iee proceedings f (communications, radar and signal processing). Vol. 131. No. 6. IET Digital Library, 1984.
[14] Lazebnik, Svetlana, Cordelia Schmid, and Jean Ponce. "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories." 2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06). Vol. 2. IEEE, 2006.
[15] Huertas, Diego Fabian Collazos, Gloria Stephany Gómez Gómez, and Andrés Marino Álvarez Meza. "Image-based animal recognition based on transfer learning." Scientia et Technica 26.03 (2021): 406-411.
[16] Sowmya, M., M. Balasubramanian, and K. Vaidehi. "Classification of Animals Using MobileNet with SVM Classifier." Computational Methods and Data Engineering: Proceedings of ICCMDE 2021. Singapore: Springer Nature Singapore, 2022. 347-358.
[17] Vidhyalatha, T., Y. Sreeram, and E. Purushotham. "Animal Intrusion Detection using Deep Learning and Transfer Learning Approaches." International Journal of Human Computations Intelligence 1.4 (2022): 51-60.
[18] Gaston, Kevin J., and Mark A. O'Neill. "Automated species identification: why not?." Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences 359.1444 (2004): 655-667.