Environmental Information Extraction Based on YOLOv5-Object Detection in Videos Collected by Camera-Collars Installed on Migratory Caribou and Black Bears in Northern Quebec

Jalila Filali[1,2], Denis Laurendeau[1,2] and Steeve D. Côté[3,4]

[1]*Department of Electrical and Computer Engineering, Faculty of Sciences and Engineering, Laval University, Quebec, Canada*

[2]*Computer Vision and Systems Laboratory (CVSL), Laval University, Quebec, Canada*

[3]*Department of Biology, Faculty of Sciences and Engineering, Laval University, Quebec, Canada*

[4]*Caribou Ungava, Centre for Northern Studies (CEN), Laval University, Quebec, Canada*

Keywords: YOLOv5 Model, Video Object Detection, Video Stabilization, Environmental Information Extraction, Data Visualization.

Abstract: With the rapid increase in the number of recorded videos, developing and exploring intelligent systems become more prominent to analyze video content. Within projects related to Sentinel North's research program*, our project involves how to analyze videos that are collected using camera collars installed on caribou (Rangifer tarandus) and black bears (Ursus americanus) living in northern Quebec. Our objective was to extract valuable environmental information such as weather, resources, and habitat where animals live. In this paper, we propose an environmental information extraction approach based on YOLOv5-Object detection in videos collected by camera collars installed on caribou and black bears in Northern Quebec. Our proposal consists, firstly, in filtering raw data and stabilizing videos to build a wildlife video dataset for deep learning training and evaluating object detection. Secondly, it focuses on solving the existing difficulties in detecting objects by adopting the YOLOv5 model to incorporate enriched features and detect objects of different sizes, and it further allows us to exploit and analyze object detection results to extract relevant information about weather, resources, and habitat of animals. Finally, it consists in visualizing object detection and statistical results by developing a GUI interface. The experimental results show that the YOLOv5m model was significantly better than the YOLOv5s model and can detect objects with different sizes. In addition, the obtained results show that our method can extract weather, habitat, and resource classes from stabilized videos, and then determine their percentage of appearance. Moreover, our proposed method can automatically provide statistics about environmental information for each stabilized video.

1 INTRODUCTION

Biodiversity has been declining at an alarming rate in recent years, mainly due to various human-driven habitat changes and the change in the earth's climate. The Living Planet Report 2022 reveals an average decline of 69% in populations of different species since 1970 (Adam et al., 2021). There is an urgent need to understand the principal causes, as well as the mechanisms of biodiversity loss. Therefore, it is fundamental to obtain timely and exact information on species richness distribution, animal behavior, wildlife resources, and animal habitats. In recent years, camera collars installed on wild animals have been widely used in wildlife surveys and to sample environmental parameters. Thus, videos and images can be collected to provide valuable information for biologists and wildlife conservation scientists. Therefore, developing and exploring intelligent systems becomes more and more prominent to analyze video content.

In recent years, deep learning models have been attracting increasing amounts of attention, due to their ability to help researchers analyze data more efficiently. Several methods based on deep learning techniques have been proposed for object detection such as animal detection (Jintasuttisak et al., 2022), wild animal facial recognition (Clapham et al., 2020), and plant disease detection ((Lakshmi and Savarimuthu, 2022) and (Sunil et al., 2021)).

*https://sentinelnorth.ulaval.ca/en/research

Filali, J., Laurendeau, D. and Côté, S.
Environmental Information Extraction Based on YOLOv5-Object Detection in Videos Collected by Camera-Collars Installed on Migratory Caribou and Black Bears in Northern Quebec.
DOI: 10.5220/0011691200003417
In *Proceedings of the 18th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2023) - Volume 5: VISAPP*, pages 660-667
ISBN: 978-989-758-634-7; ISSN: 2184-4321

In (Norouzzadeh et al., 2018), a deep convolutional neural network is used to automatically identify, count, and describe wildlife images with high classification accuracy. Recently, You Only Look Once (YOLO) model (Redmon et al., 2016), the first single-stage detector in deep learning has been widely applied for object detection.

Nowadays, several object detection models have been proposed in ecology and biology. However, object detection for extracting valuable information from videos is still a challenging task and most of the proposed object detection methods do not take the whole image content into consideration to extract information about weather, about weather, resources, and habitat where animals live. Moreover, there are few wildlife video datasets for deep-learning training. Indeed, it becomes difficult to extract good-quality images from videos that are collected by cameras carried by animals. This is due to the movement of animals that influences the video quality (unstabilized videos, blurred images, visible distortion, etc.). In this paper, we report our attempts to address the above issues, as follows: First, we constructed a wildlife video dataset for deep learning training and evaluating object detection. Second, we focused on solving the existing difficulties in detecting objects by adopting the YOLOv5 model to incorporate enriched features and detect objects of different sizes, and we further exploited and analyzed the object detection results to extract relevant information about weather, resources, and habitats that are found in the environment in which caribou and black bears live. Finally, we are interested in visualizing object detection and statistics results by developing a GUI interface.

The main contribution of our work is summarized as follows: First, we propose four video stabilization methods based on motion compensation with different parameter combinations and we compare their performances to determine which method provides a better stabilization quality. Then, we study the relevance of stabilized videos for object detection. To this end, we propose a relevance score for object detection that can predict if a given stabilized video contains interesting objects that can be used for extracting information. Second, we adopt the YOLOv5 model to detect objects in stabilized videos. Both YOLOv5s and YOLOv5m models are tested with different parameters using a large dataset, training results and object detection performance are evaluated. Finally, our proposed method can automatically extract weather, habitat, and resource information from a given stabilized video and identify the percentage of appearance of each class in the video.

This paper is organized in the following way. Section 2 presents an overview of the related research. Our proposed method is detailed in Section 3. Experimental results are presented and discussed in Section 4. Finally, some final remarks and directions for future work are included in Section 5.

2 RELATED WORKS

2.1 Video Stabilization

Video stabilization (VS) consists in transforming a video corrupted by undesired camera motions into a stabilized video to improve its quality by removing unwanted camera shakes and jitters. In the literature, several methods of VS have been developed to produce a better video quality and a coherent video stream. The existing VS approaches can be generally categorized into classical VS methods and learning-based methods(Shi et al., 2022). Classical video stabilization algorithms focus on estimating camera motion, correcting camera path and stabilizing the video (Guilluy et al., 2021). Learning-based video stabilization methods consist in stabilizing videos using learning and deep-learning models (Shi et al., 2022).

Generally, using classical VS methods, the video stabilization process includes two main phases, namely 1) Motion analysis and modeling and 2) Motion correction and video stabilization. Motion analysis and modeling go through three steps: Motion estimation; 2) Motion outlier removal; and 3) Motion modeling. The aim of the first step is to estimate the camera motion from the original video. These estimated movements can be divided into two types of motion: camera motion and object motion that appears in the scene. To that end, the second step, namely motion outlier removal, is carried out to remove outliers and select only those resulting from the motion of the camera. These movements are then used for camera motion modeling. This step allows to model the camera motion (Guilluy et al., 2021) and compute the global transformation (Kulkarni et al., 2016).

The motion correction and video stabilization phase includes two steps: motion smoothing and video synthesis. In this phase, parameters of the previous phase are sent to the motion smoothing module, the purpose of which is to perform camera motion correction. This step allows to apply a low-pass filter to remove the high-frequency distortion and smooth the camera movements. Once the camera motion has been corrected, the new camera movements are applied back to the original video and a new video using

the smoothed camera movements is reconstructed as a stabilized video within the video synthesis step.

2.2 Video Object Detection

Advances in artificial intelligence and computer vision have enabled the creation of efficient systems to analyze video content and extract semantic information. In this context, object detection is one of the most important and challenging problems in computer vision. It entails identifying and localizing all instances of an object in input images or videos. In the literature, numerous approaches have been proposed to solve the object detection problem. In general, these approaches can be divided into two categories: object detection methods based on handcrafted features and deep learning-based object detection methods.

Object detection methods based on handcrafted features, also called traditional methods, revolve around extracting features from images that are used for object detection. Using these methods, object detection models were built as a set of hand-crafted feature extractors such as Viola Jones Detectors that have incorporated feature selection and techniques to increase the detection speed (Viola and Jones, 2004). In addition, several local feature descriptors were applied to address the problem of scale variations and rotations. In this context, various handcrafted feature descriptors have been proposed to extract relevant information from images. Among these descriptors, we distinguish scale-invariant feature transform (SIFT) (Lowe, 2004), speeded-up robust features (SURF) (Bay et al., 2006) and Histogram of Oriented Gradients (HOG) (Dalal and Triggs, 2005). These methods have used a constructed feature pyramid with a fixed sliding window to detect objects. Despite that the traditional object detection methods cannot meet the requirements for video data analytics, they have provided a strong foundation for future video object detection systems. Deep learning-based object detection methods can be divided into two categories: two-stage object detection methods and single-stage object detection methods.

The most representative two-stage object detection approaches are Regions-based Convolutional Neural Networks(R-CNN) (Girshick et al., 2014). The goal of R-CNN is to extract a set of object proposals using Selective Search (Uijlings et al., 2013), and then each proposal is warped and propagated through the convolutional layers. Finally, a feature vector is extracted from each proposal and used as an input of linear SVM (Support Vector Machine) classifiers to compute confidence scores. Once the class has been recognized, the algorithm predicts its bounding box using a trained bounding-box regressor.

The two-stage object detection approaches solve object detection as a classification problem where the network classifies image content as either object or background. However, single-stage object detection approaches solve it as a regression problem that is performed without using pre-generated region proposals, the main goal being to predict the image pixels as objects and their bounding box attributes.

In the literature, several single-stage object detection methods have been proposed. In 2015, You Only Look Once (YOLO) was proposed by (Redmon et al., 2016) as the first single-stage detector in deep learning. YOLO was inspired by the GoogLeNet architecture for image classification (Szegedy et al., 2015). The input image is divided into a S x S grid where each cell of the grid is responsible for object detection. This network predicts bounding boxes with their confidence scores for each grid cell simultaneously. In recent years, the YOLO network has been improved and it was a milestone in object detection due to its efficiency in real-time with better accuracy. The second version YOLOv2 (Redmon and Farhadi, 2017) incorporated many techniques to improve speed and precision. In the original YOLO, attributes of predicted boxes were generated by fully connected layers. In YOLOv2, the fully connected layers are removed, this version used anchor boxes to generate offsets as well as predicted boxes. Classes and objectness are predicted for every anchor box. YOLOv3 (Redmon and Farhadi, 2018) proposed a larger and robust feature extractor network called Darknet-53. YOLO has further been improved in the following versions, including YOLOv5 (Jocher et al., 2022) which is proposed by Glenn Jocher in 2020, and YOLOv7 (Wang et al., 2022).

3 PROPOSED APPROACH

In this section, we describe the architecture of our proposed approach and detail the different phases and their components. As depicted in Figure 1, the proposed approach is composed of four main phases: (1) Data preprocessing, (2) Video object detection, (3) Environmental information extraction and (4) Results Visualization. The different components are detailed below.

3.1 Data Preprocessing

In our project, camera collars are used as tools for collecting videos. These videos are gathered from cam-

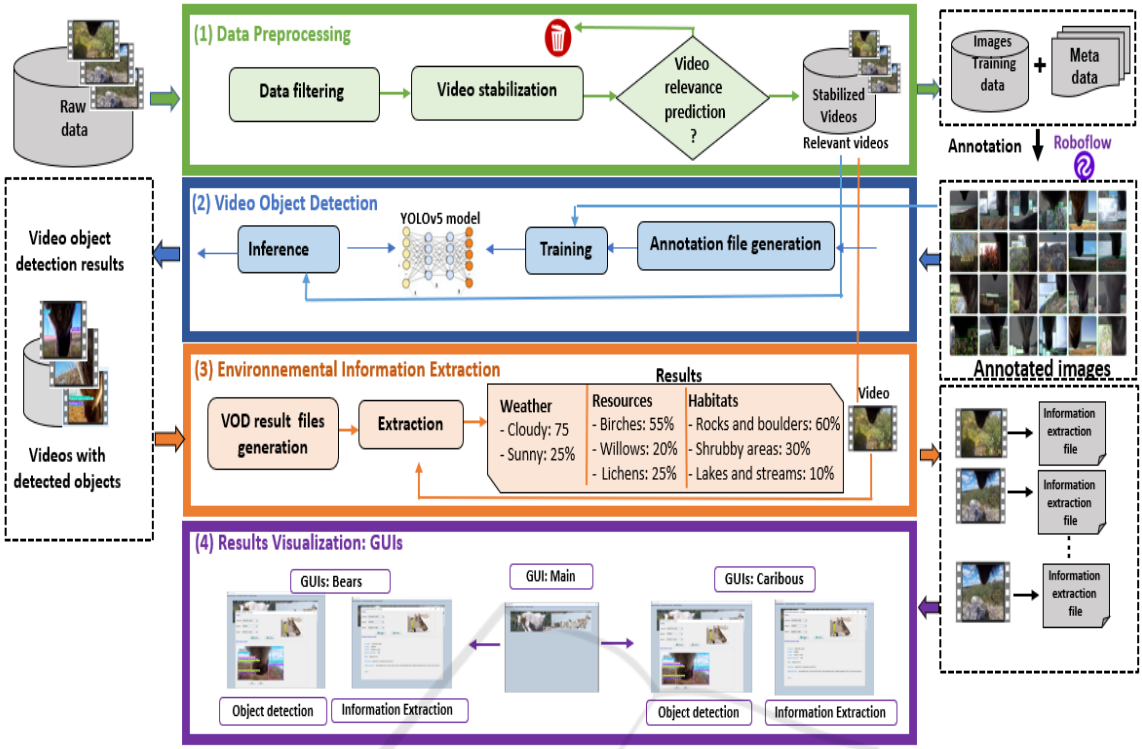

Figure 1: Architecture of the proposed approach.

eras mounted on collars that are carried by caribou and black bears. Due to the movement of the animals, videos are often low quality (unstabilized videos, blurred images, visible distortion, useless data, etc.). To this end, video processing needs data filtering and video stabilization as a prerequisite for object detection. The data preprocessing phase includes three components, namely: data filtering component, video stabilization component, and video relevance prediction component.

3.1.1 Data Filtering

It is observed that several videos contain blurry and noisy frames due to the undesired motion of cameras carried by the animal. In addition, other videos suffer from a lot of frames that do not contain any information related to the animal's environment. Data filtering is an important step to clean our dataset to be provided as input for the next process. The data filtering process consists in rejecting videos that contain only dark frames, and videos that contain blurred images which do not represent any information about the environment of animals.

3.1.2 Video Stabilization

In our work, we focus on classical video stabilization methods. To stabilize videos, we proposed four stabilization methods: 1) Video stabilization based on movement compensation (translation and rotation) without limit; 2) Video stabilization based only on camera motion (translation) with limited rotation; 3) Video stabilization based only on rotation with limited camera motion, and 4) Video stabilization based on sequential combined method: method 2 + method 3. The main idea is to perform video stabilization by incorporating motion compensation with different parameter combinations and then to compare their performances to determine which method can provide a better stabilization quality. The first method consists in stabilizing video based on compensation of translation and rotation movement without limit, the second one allows to stabilize only camera motion with limited rotation. As opposed to the second method, the third method involves only rotation to compensate motion to obtain the stabilized output video. The last method consists in performing stabilization in two stages, by using a sequential combined method which allows applying the second method, and then the obtained stabilized videos are used to perform a second stabilization that is based on the third method. The proposed stabilization methods have many parameters that influence stabilization performance. The parameter combinations are presented in Section 4.2.1.

Table 1: Video stabilization results.

Methods	Shakiness	Accuracy	Smoothing	Optalgo	Interpol	Maxshift	Maxangl	PSNR(dB)
Method 1	5	8	30	gauss	linear	no limit	no limit	35,40
Method 2	10	15	15	avg	bilinear	no limit	0	40,95
Method 3	10	15	15	avg	bilinear	0	no limit	42,07
Method 4	10	15	15	avg	bilinear	nolimit	0	**47,55**
	10	15	15	avg	bilinear	0	no limit	

3.1.3 Video Relevance Prediction for Object Detection

Once raw videos have been filtered and stabilized, a study should be carried out to predict if the stabilized videos are relevant for object detection. The aim is to predict if a given stabilized video contains interesting objects that will be used for extracting information about animal environment. To achieve this goal, we propose a relevance score for object detection.

Let us present the following definitions to explain our proposed relevance score:

- V is a given stabilized video.
- $F = F_1, F_2, ..., F_N$ is the set of frames extracted from a given video V,
- $P = P_{i=1...M} F_j$ is the set of patches (a patch presents a group of pixels in the frame and each frame is divided into small patches) extracted from the frame F_j.
- P_t is the target patch of the V.
- STP_{F_j} is the number of similar patches to the target patch P_t of the frame F_j.
- NP_{F_j} number of patches processed for the frame F_j.

Given the stabilized video V and the target patch P_t, the aim is to determine a relevance score $Rs_{det}(V)$ for each video V to predict if it is relevant for object detection. Our main idea is, firstly, to compute a weight for each frame $w(F_j)$ depending on the number of similar patches to target patch STP_{F_j}, and then to weigh the similarity between every two adjacent frames of the video using the weight $w(F_j)$, and finally to compute the weighted average of all video frames. We define the following functions to compute the proposed relevance score for object detection:

- $W(F_j) = 1 - \frac{STP_{F_j}}{NP_{F_j}}$
  Where : $W(F_j)$ is the weight of the frame F_j

- $Sim(F_j, F_{j+1}) = \frac{1}{1+d(F_j,F_{j+1})}$
  Where : $Sim(F_j, F_{j+1})$ is the similarity between the frames F_j and F_{j+1} and $d(F_j, F_{j+1})$ is the Euclidean distance between frames.

These two functions are used for computing the proposed relevance score $Rs_{det}(V)$:

$$Rs_{det}(V) = \frac{\sum_{j=1}^{|N|} w(F_j) * Sim(F_j, F_{j+1})}{\sum_{j=1}^{|N|} w(F_j)} \quad (1)$$

3.2 Video Object Detection

To detect objects from stabilized videos, we adopted feature enriched object detector based on the YOLOv5 model. Our choice of YOLOv5 network is based on its detection speed and accuracy, which incorporates better parameter structure into the backbone. Its detection speed and accuracy on COCO datasets are better than previous YOLOv4 and YOLOv3 models. Also, YOLOv5 incorporated various optimization techniques like auto-learning bounding box anchors, data augmentation, and the cross-stage partial network (CSPNet). In addition, the head of Yolov5 generates 3 different sizes (18×18, 36×36, 72×72) of feature maps to perform multi-scale prediction that allows to handle small, medium, and large objects. In our project, many resources and habitats of caribou and black bears are characterized by their small size like "lichens" and "rocks". These objects can be transformed into medium and large objects depending on the camera position. To this end, multi-scale detection ensures that the model can follow size changes in the process of object detection.

The YOLOv5 model consists of an input layer of 640 x 640 image, Backbone, Neck, and Head. The backbone consists of Focus, Conv, C3 (CSPNet Bottleneck with three convolutions), and Spatial Pyramid Pooling (SPP) modules. The Focus module divides the mosaic input image horizontally and vertically and then stitches it together. The main purpose of the Focus layer is to reduce layers, parameters, and FLOPS (Floating-point Operations per Second). Conv is the convolution unit of YOLOv5, it performs dimensional convolution, regularization, and activation operations. C3 contains 3 Convs and some bottlenecks. These BottleneckCSPs are used to reduce the number of calculations and increase the speed of inference. The spatial pyramid pooling (SPP) layer performs three different sizes of maximum pooling operations on the input, and the output result is spliced

into Concat. Upsample modules are used by the 11 and 15 layers of the Neck Network to expand features. The head of YOLOv5 has three feature detection scales that allow to achieve feature detection of different sizes.

3.3 Environmental Information Extraction and Results Visualization

Once objects are detected, a video object detection result file is generated for each stabilized video, which contains details about which camera is used, the number of object detection, and the detected classes of each category (weather, resources, and habitats). VOD result files are then used to determine the appearance percentage of each object that is detected in the video (cf. Figure 1). The generated result files are then exploited to visualize object detection and environmental information extraction results with a GUI interface.

4 EXPERIMENTATION AND RESULTS ANALYSIS

4.1 Experimental Setup

4.1.1 Dataset and Image Annotation

In our work we used about 22597 videos which are collected by 4 cameras. To prepare data for training models, we annotated about 1177 images using the Roboflow framework [1]. A semantic vocabulary is used to describe the content of videos. It can be divided into 3 categories: weather, habitats, and resources which have their own classes and are used as annotation labels and were respectively defined as follows: **Weather**: sunny, cloudy, rain, fog, snow; **Habitats:** snow on the ground, boreal forest, tundra meadow, shrubby areas, wet meadow, lakes and streams, rocks and boulders and **Resources:** lichens, birches, willows, grasses and sedges, broad-leaved herbaceous, mushrooms, berries. The images were divided into a training set, a validation set, and a test set. The number of images in the training set, validation set and test set was 937 (80%), 120(10%) and 120(10%), respectively.

[1]https://roboflow.com/

4.1.2 Evaluation Metrics

Peak Signal-to-Noise Ratio (PSNR) is used to evaluate the quality of stabilized videos. PSNR computed between the original video and the stabilized video is defined as:

$$PSNR = 10log_{10}\frac{Max_I{}^2}{MSE} \quad (2)$$

Where MSE measures the Mean-Square-Error between the original and stabilized video frames and MAX_I is the maximum pixel value of an image. Greater PSNR values indicate better quality video stabilization.

To evaluate object detection performance, we used precision and recall, which are the most popular metrics for object detection evaluation.

4.2 Evaluation Results and Analysis

4.2.1 Video Stabilization

Table 1 shows the stabilization results of each method in terms of PSNR depending on the stabilization parameters using 1500 videos. On performing a detailed analysis of each of the above mentioned methods, it was found that Method 4 has a higher PSNR (47,55) value when compared to the other methods. The higher the PSNR, the better video stabilization. Thus, the stabilized output videos obtained by the 4th method are better than those stabilized using the other methods. This explains that performing stabilization with rotation and translation movement compensation in two stages improves the video stabilization quality.

4.2.2 Video Relevance Prediction

To predict if the stabilized videos contain relevant objects that can be used for object detection and extracting information about weather, resources, and habitats, firstly, a patch set of each video is extracted, and then the proposed relevance score of each video is computed using the proposed functions.

- **Patch Extraction.** Table 2 shows the patch extraction results for 4615 black bear videos and 4590 caribou videos.
- **Relevance Score Analysis and Prediction.** Once patches are extracted, the relevance score $Rs_{det}(V)$ for each video is computed. To predict if videos are relevant for object detection, we used a prediction threshold of 0.5. Therefore, if the relevance score is larger than the threshold, videos are predicted as relevant for object detection, if not, they are predicted as irrelevant. Table

Table 2: Patch extraction results.

Videos	Avg(Frames\vid)	Patches\frames	Avg(patches\vid)	Rejected_patches
4615 (black bears)	420	40	16800	2100
4590 (caribou)	445	32	14240	5785

Table 3: Prediction analysis.

Cam	Videos	R.PV($Rs_{det} \geq 0.5$)	I.PV($Rs_{det} \leq 0.5$)	R.PV_STP	I.PV_STP
Cam1424_ID37585	4450	3627	823	14	27
Cam1435_ID37584	4128	2710	1418	11	31
Collar21226_cam06	3678	1178	2500	12	29
Collar21227_cam07	4023	2245	1778	13	22

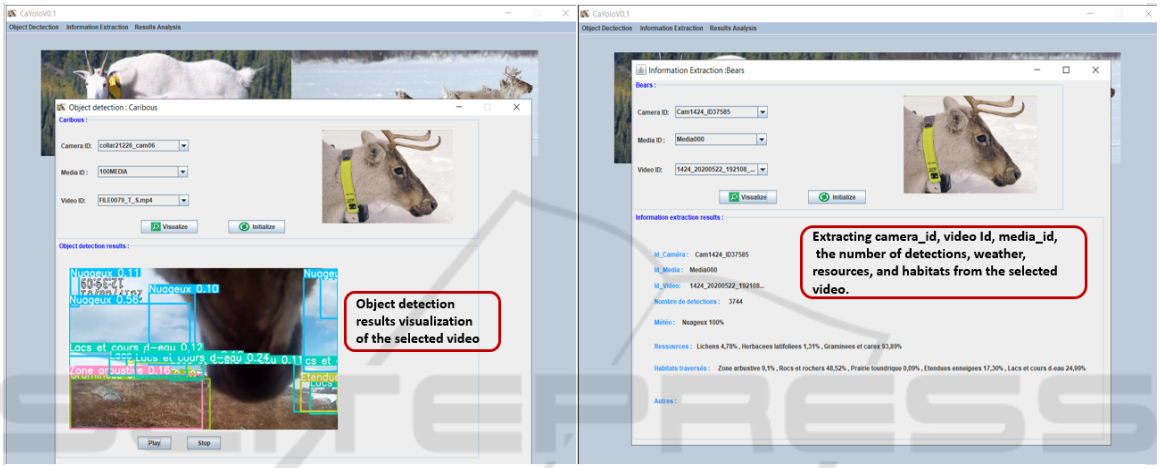

Figure 2: Examples of two interfaces for object detection and information extraction results visualization: the first one shows the visualization of object detection results of a selected caribou video, and the second one shows the visualization of environmental information extraction results of a selected black bear video.

3 shows the prediction analysis for each camera where *R.PV* and *I.PV* is the number of videos that are predicted as relevant and irrelevant, respectively, for object detection. *R.PV_AVG_STP* and *I.PV_AVG_STP* is the average of similar patches to the target patch per frame for *R.PV* and *I.PV* videos, respectively. The obtained results show that the black bear videos are better than the caribou videos in terms of pertinence for object detection.

4.2.3 Object Detection and Information Extraction Results

The experimental model was trained using Ultralytics YOLOv5. We performed 4 training runs using YOLOV5m and YOLOV5s with 300 epochs. Both YOLOV5s et YOLOV5m were implemented using 16 and 32 batch sizes. The comparison results of object detection in terms of precision and recall are shown in Table 4. It is found that the YOLOv5m-B32E300 model outperforms with 0.81 precision. Therefore,

Table 4: Comparisons of different runs.

Run	imges	P	R
YOLOv5s-B16E300	100	0.61	0.41
YOLOv5s-B32E300	100	0.69	0.42
YOLOv5m-B16E300	100	0,78	0.44
YOLOv5m-B32E300	100	**0.81**	0.46

the YOLOv5m model with batch size adjusted to 32 is highly capable of detecting objects with different sizes, so it is used for object detection and environmental information extraction. Examples of two interfaces for object detection and information extraction results visualization are detailed in Figure 2.

5 CONCLUSIONS

In this paper, an environmental information extraction method based on YOLOv5-object detection in videos was proposed. It relies on analyzing videos collected by camera collars fitted on caribou and black bears

in northern Quebec. First, videos are filtered, stabilized, and then, a relevance score for object detection is proposed and computed for each stabilized video to predict if it is relevant for object detection. Second, the YOLOv5 model is adopted to incorporate enriched features and detect small, medium, and large objects. Object detection results are then exploited to extract relevant information about weather, resources, and habitats found in the environment in which caribou and black bears live. Finally, the environmental information is analyzed and statistical results are visualized for each stabilized video. In this work, we have conducted an experimental study where we focused on evaluating each phase of our proposed approach. It is worth to note that the proposed stabilization method, based on motion compensation with different parameter combinations, can improve the quality of the videos. Also, the YOLOv5m model was significantly better than the YOLOv5s model and can detect small, medium, and large objects. Moreover, the obtained results show that our method can extract weather, habitat, and resource classes and then determine their percentage of appearance in videos. In future research, the network model structure will be improved to analyze animal behavior using a wildlife dataset.

REFERENCES

Adam, M., Tomášek, P., Lehejček, J., Trojan, J., and Jůnek, T. (2021). The role of citizen science and deep learning in camera trapping. *Sustainability*, 13(18):10287.

Bay, H., Tuytelaars, T., and Gool, L. V. (2006). Surf: Speeded up robust features. In *European conference on computer vision*, pages 404–417. Springer.

Clapham, M., Miller, E., Nguyen, M., and Darimont, C. T. (2020). Automated facial recognition for wildlife that lack unique markings: A deep learning approach for brown bears. *Ecology and evolution*, 10(23):12883–12892.

Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, pages 886–893. Ieee.

Girshick, R., Donahue, J., Darrell, T., and Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587.

Guilluy, W., Oudre, L., and Beghdadi, A. (2021). Video stabilization: Overview, challenges and perspectives. *Signal Processing: Image Communication*, 90:116015.

Jintasuttisak, T., Leonce, A., Sher Shah, M., Khafaga, T., Simkins, G., and Edirisinghe, E. (2022). Deep learning based animal detection and tracking in drone video footage. In *Proceedings of the 8th International Conference on Computing and Artificial Intelligence*, pages 425–431.

Jocher, G. et al. (2022). ultralytics/yolov5: v6. 2-yolov5 classification models, apple m1, reproducibility, clearml and deci. ai integrations. *Zenodo. org*.

Kulkarni, S., Bormane, D., and Nalbalwar, S. (2016). Stabilization of jittery videos using feature point matching technique. In *International Conference on Communication and Signal Processing 2016 (ICCASP 2016)*, pages 708–717. Atlantis Press.

Lakshmi, R. K. and Savarimuthu, N. (2022). Pldd—a deep learning-based plant leaf disease detection. *IEEE Consumer Electronics Magazine*, 11(3):44–49.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110.

Norouzzadeh, M. S., Nguyen, A., Kosmala, M., Swanson, A., Palmer, M. S., Packer, C., and Clune, J. (2018). Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning. *Proceedings of the National Academy of Sciences*, 115(25):E5716–E5725.

Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788.

Redmon, J. and Farhadi, A. (2017). Yolo9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7263–7271.

Redmon, J. and Farhadi, A. (2018). Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*.

Shi, Z., Shi, F., Lai, W.-S., Liang, C.-K., and Liang, Y. (2022). Deep online fused video stabilization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1250–1258.

Sunil, C., Jaidhar, C., and Patil, N. (2021). Cardamom plant disease detection approach using efficientnetv2. *IEEE Access*, 10:789–804.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9.

Uijlings, J. R., Van De Sande, K. E., Gevers, T., and Smeulders, A. W. (2013). Selective search for object recognition. *International journal of computer vision*, 104(2):154–171.

Viola, P. and Jones, M. J. (2004). Robust real-time face detection. *International journal of computer vision*, 57(2):137–154.

Wang, C.-Y., Bochkovskiy, A., and Liao, H.-Y. M. (2022). Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *arXiv preprint arXiv:2207.02696*.