Semantic Annotation of Videos Based on Mask RCNN for a Study of Animal Behavior

Nourelhouda Hammouda[1(✉)], Mariem Mahfoudh[1,2], and Mohamed Cherif[3]

[1] University of Kairouan, Kairouan, Tunisia
hamouda.nourelhouda@gmail.com
[2] MIRACL Laboratory, University of Sfax, Route de Tunis Km 10, B.P. 242, 3021 Sfax, Tunisia
[3] INRAT Animal and Forage Production Laboratory, Ariana, Tunisia

Abstract. Detection and tracking of object video are of great interest in various fields like security, traffic, and public places... In this article, we are interested in agriculture and animal husbandry. The majority of existing video object tracking applications aim to extract the object's trajectory based on the detection results and then correlate the obtained coordinates in most frames. But to build knowledge about animal behavior in open pasture, and how long it takes for each behavior, we need the coordinates of each object in all the video's frames. Therefore, in our paper, we aim in the first step to extract all the information about each animal (sheep) from the video. Our research is not limited to knowing behaviors only, but we also seek in the second step to collect the information extracted using specific rules to build knowledge about the effects of animal behavior on themselves, on the pastures, and to know the status of the pastures. This knowledge subsequently contributes to making the best decisions to preserve the vegetation cover and the quality of animal production. We use the Mask R-CNN detector and MS COCO dataset to detect and extract the location of the animal in each frame. We use the Hungarian algorithm to associate similar objects in all video frames. Then we correct the detection and association mistakes. Our approach was able to achieve 100% tracking of all sheep if the sheep not moving very fast. Finally, we use ontologies OWL to represent and extract Knowledge and we express in SWRL the semantic rules which help us to study animals behavior.

Keywords: Video annotation · Deep learning · Mask RCNN · Semantic annotation · Ontology · Animal behavior · Making decision · SWRL

1 Introduction

The vegetation cover of the natural pastures is characterized by the diversity and richness of its components that can improve the quality of animal production. This subject has received great attention from the National Institute of Agricultural Research in Tunisia (INRAT), to take advantage of these characteristics

G. Memmi et al. (Eds.): KSEM 2022, LNAI 13370, pp. 391–402, 2022.
https://doi.org/10.1007/978-3-031-10989-8_31

while maintaining its permanent presence. Therefore, they suggest studying the behavior of animals in pastures with a smartly and accurately method that helps them in making good use of this wealth. On the one hand, the study of the behaviors helps to know the type of plant compatible with each animals breed and the effect of these behaviors on itself. On the other hand, animal behavior helps us to know the effect of grazing on the pastures. Thus, we can control it to preserve the vegetation cover from desertification.

Many studies have already started using automatic surveillance techniques to monitor animal behavior and to analyze human behavior. [9] and [5] surveyed lots of research contributions in this domain. They mentioned that improving the accuracy and quality of the detector played a very important role in correctly tracking objects in the video. The most notable are deep learning-based detection methods such as Faster R-CNN, Mask R-CNN, YOLO, SSD... In this context, [1] carried out a full survey of most work interested to solve the mission of tracking multiple MOT (Multiple Object Tracking) objects on single-camera videos, these algorithms consist of four main steps, the detection step is the first step.

Annotation of visual data is important as it provides ground truth labels of real-world objects, scenes, and events... But the manual annotation is a complex, time-consuming, and laborious task that can only be performed by one expert to ensure the accuracy of the annotation and use the same criteria. Recently, [15] noted that ontology is of paramount importance in facilitating semantic communication between different metadata to provide a semantic description of images. We also find that the ontologies and their backing technologies like OWL, construct more complex semantic classes by combining and intersecting existing concepts [14].

In our paper, we are interested in extracting knowledge from a video with a high-resolution and fixed camera. This phase requires the detection and recognition of the animals (sheep) based on Mask RCNN detector, using MS COCO dataset. Then the tracking of each animal throughout the video sequence using Hungarian algorithm to facilitate the association process. To determine the behavior of each animal later.

Our paper is organized as follows: in the next section, we will describe related work on detection, tracking, extraction, and analysis knowledge. In Sect. 3, we will explain our approach, the methods in which we build our work, and the contributions we have made to improve the tracking process and to study the knowledge extracted. Then, we will evaluate and discuss our approach results.

2 Related Works

This section reviews previous techniques for detecting and tracking multiple objects in video and the contributions of ontology to the study of behavior and decision support.

2.1 Video Annotation

"A video is the result of a sequence of frames displayed with a sufficiently fast frequency" [12]. Detecting and recognizing objects in all video frames, then associating each object with all its images throughout the video poses many challenges (occlusion, shadow, no-rigid object deformation...) [19].

Recently, deeper CNNs (DCNNs) have led to unprecedented improvements in the detection of more general categories of objects [10,20]. CNN's successful applications to image classification have been transferred to object detection, resulting in a region-based CNN detector (RCNN) [2]. Since then, much object detection research has drawn on the rapidly evolving area of RCNN work. Since 2013, in an attempt to improve performance, many detectors have been proposed. Additionally, [9] presented a survey on the properties and performance of infrastructures for generic object detection.

They presented two main categories of detectors: Region-based (Two-Stage Framework) and Unified Pipeline (One-Stage Pipeline). In the first category, there are RCNN, which were among the first to explore CNN for generic object detection and developed RCNN, This approach has seen several improvements, including the SPPNet, Fast RCNN, Faster RCNN, RFCN until it reached then proposed the Mask RCNN to tackle the segmentation of object instances at the pixel level by extending Faster RCNN. Mask RCNN adds a branch that generates a binary mask for each RoI. The new branch is a fully convolutional network (FCN) located on top of a CNN feature map. To avoid misalignment caused by the original RoI Pooling (RoIPool) layer, a RoIAlign layer was proposed to preserve spatial correspondence at the pixel level. With a ResNeXt101-FPN backbone [8], Mask RCNN achieved the best results in terms of COCO object instance segmentation and bounding object detection. And the second category, among the algorithms, includes the YOLOv2 and YOLO9000: proposed YOLOv2, an improved version of YOLO. It reached the state of the art in standard detection tasks such as PASCAL VOC and MS COCO. SSD (Single-shot detector): To preserve real-time speed without sacrificing excessive detection accuracy. SSD effectively combines RPN ideas in Faster RCNN, YOLO, and multi-scale CONV functions to achieve fast detection speed while maintaining high detection quality.

Detection has been used for several purposes, the most important is to track objects which have made significant progress in the last decade. [18] noted recently, that some researchers used online and offline CNN method processing to segment and track at one time and it showed fast and accurate tracking results and segmentation. In addition, a full investigation has been done by [1], the authors reported that online algorithms are too slow compared to batch tracking algorithms, especially when using algorithms that often require a lot of computation like deep learning algorithms. Most of the MOT algorithms share some of their steps. We mention, for example, some of the proposed approaches: [21] suggested a tracking method consisting of four stages: Detection, extraction of appearance characteristics, and the association then tracking of moving

objects. And the approach of [6] consists of only two phases: online tracker and then association.

We noticed in most of these works that they use the appearance features, applied and evaluated on humans only, also they use the center of the detection frame or a small part of the object to track the movement of the object. [1] reported that methods that do not exploit appearance are less efficient. They also showed that the most efficient detectors currently are the Faster R-CNN and its variants. [10] points out that the tracking-by-detection model is widely used, but it still faces some difficulties, and there are several reform attempts recently.

2.2 Behavior Study

We need now to analyze the data extracted from the first stage like the duration of each activity, the indicator of each activity duration, the season during which these activities take place, etc.

According to [16], ontology is a discipline of philosophy, it is the study of being as an entity, of what is? what type, what structure? their properties? Among these works, we cite: [4] went further than the formal definition of the various management behavior specifications integrated into the management information definitions, they focused on the definition of the rules of behavior in the management information with SWRL, a rules language defined to complete the OWL functionality. [3] used the ontology to present a semantic classification method based on objects in high-resolution satellite imagery. [13] also used the ontology to model human behavior. [17] used it to build a Semantic Ship Behavior Model (SMSB) to analyze potential ship behaviors.

3 Proposed Methodology

Our study is composed of two phases, as shown in the Fig. 1. The first consists of video annotation; extracting from the video, all information about each target object. The second step is to make the annotations semantic by studying and analyzing the knowledge constructed based on the data extracted. In phase 1, we aim to extract information from the video that could help us to analyze the behavior of animals. So in the first step, we will detect all the animals and then track them throughout the video. To track the animals, we chose the method closest to our needs because most of the existing solutions are applied to objects that can be easily distinguished, such as pedestrians and cars... However, in our case, the sheep are almost very similar. And in each step, we will try to identify all the errors and possible gaps that could affect the quality of the tracking and check all the results before recording. In phase 2, we convert all the given rules to ontology models to help us make decisions when defining and analyzing behavior.

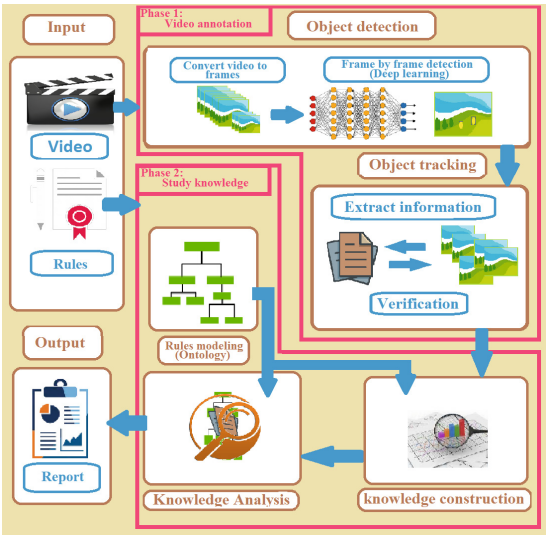

Fig. 1. Our proposed approach.

3.1 Video Annotation

Detection Step. To choose the appropriate detector, we make a comparison between the two detectors, SSD and Mask R-CNN. We test the accuracy and reliability of object identification according to each detector. We used six videos: two videos each contain a single animal, one video contains six animals, two videos contain more than six animals, and one video contains a very large number of animals.

Figure 2(a) shows the Mask RCNN detector result. It could detect all animals even if there is an occlusion. But, he rarely fails to spot certain animals. It can detect a large number of animals. From time to time, it detects part of the animal and/or part of the background as a single object or as two objects. Moreover, sometimes he considers the same animal, whether the whole body or part of it, as two animals. These faults appear frequently, especially when animals are very close to the camera or their colors are similar to the color of grass.

Figure 2(b) shows one of the SSD detector results. It could detect the animal most of the time, even if there is an occlusion when we have a little number of sheep. While it can only detect 4 to 7 animals in other videos. In a few moments, it detects the same animal as two animals. But, for several moments, he was unable to detect anything despite the 'Sheep' being very conspicuous. From time to time, the size of the box is slightly larger compared to the detected object. Also, it was remarkable that some of the chests were in the wrong location.

Based on these tests, we can easily compare the performance of each detector. Using Table 3, we can make a decision about which detector will be used in our project. We have in this table V_i with i = [1..6]: Video number, nb: Number of sheep per video, DON: Detected Objects Number,

(a) (b)

Fig. 2. Extracts of videos, objects detected by Mask RCNN (a) and by SSD (b).

DON/Video	OTD		OTD_IP		OTD_IN		$NTOD_T$	
Detector	MRCNN	SSD	MRCNN	SSD	MRCNN	SSD	MRCNN	SSD
$v_1(nb = 1)$	1-2	0-2	1	0-1	0-1	0-1	1	0
$v_2(nb = 1)$	1-2	0-2	1	0-1	0-1	0-1	1	0
$v_3(nb = 6)$	6-2	5-8	4-3	3-6	1-0	1-2	0	0
$v_4(nb \gg 15)$	18-14	3-5	14-3	4-3	3	1	0	1-2
$v_5(nb \gg 15)$	24-17	1-3	20-14	1-3	3-4	1	0	0-2
$v_6(nb \gg 20)$	20-7	0-4	19-1	4-0	6-1	0	0	0

Fig. 3. Comparison between the ability of each detection device to detect and identify animals.

Videos	V_1		V_2		V_3		V_4		V_5		V_6	
Errors	Mask	SSD	Mask	SSD	Mask	SSD	Mask	SSD	Mask	SSD	Mask	SSD
$Error_1$	No	Yes	No	Yes	Yes	Yes	No	Yes	No	Yes	No	Yes
$Error_2$	Yes	Yes	Yes	No	No	No	No	Yes	No	Yes	No	No
$Error_3$	Yes	Yes	Yes	Yes	Yes	Yes	No	No	No	No	No	No
$Error_4$	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No

Fig. 4. Errors occurred in each test.

DTO: Detected Target Objects number, DTO_IP: number of Detected Target Objects Positively Identified, DTO_IN: number of Detected Target Objects Negatively Identified, $NTOD_T$: number of Non-Target Objects Detected as Target objects. The numbers are given here present in all moments of the videos.

Video number i, Yes: the $error_j$ exists, No: the $error_j$ does not exist step, j = [1..4].

We notice, using this table, that the Mask RCNN detector is the least likely to make the errors in the V4 and V5 videos, while they are equal in the V6 video. But in the V6 video, the only error that the SSD detector makes is $error_1$, but we can't rely on his performance, because he could only detect four of more than twenty creatures.

Four types of errors we observed during these tests that they are: The failure to detect all the animals throughout the video ($error_1$), to detect non-target objects as objects targets ($error_2$), detect a single target object as two different objects ($error_3$) and poorly recognizing objects ($error_4$). We simplify the existence of errors in all videos with Table 4. We have in this table V_i with i = [1..6]:

After these results, we choose the Mask RCNN detector to use in our project. Although the results are acceptable under conditions similar to those of the V4 and V5 videos, the possibility of errors remains probable. Certainly, this will affect the proper tracking of each object. Therefore, we will do our best to avoid as many errors as possible and ensure good tracking.

Tracking Step. After important searches, we found that almost all searches depend on the box object detected, their center or the contour of mask object detected ... etc. to associate the pairs of the objects. But in our case, these methods can make false associations. As we show in Fig. 5(a), we cannot specify the

true center of each object during an occlusion. Therefore, we propose another method for a good selection of the center that gives the true object position.

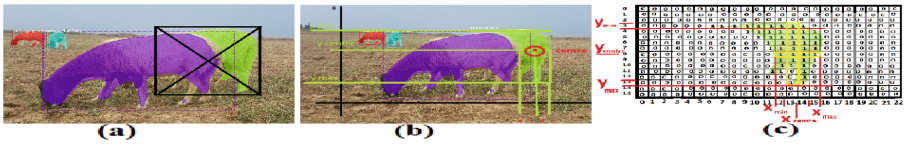

Fig. 5. Solution proposed to improve position selection.

The Mask R-CNN detector gives us a matrix of 0 and 1, as shown Fig. 5(c), for each object mask, it has the same image Fig. 5(b) dimension. Our proposal is:

$$Y_{centre} = (Y_{min} + Y_{max})/2$$
$$X_{centre} = (X_{min} + X_{max})/2(\text{of line } Y_{centre})$$

y_{min} is the first-row index and y_{max} is the last row index, which contain one or more columns with "1". X_{min} is the first column index and X_{max} the last column index in the line y_{center} which have "1". With this method, we can associate the centers of objects in the image (i-1) and in the image (i), using the Hungarian algorithm. The next position of the object S is the position of the center (i) closest to its center (i-1).

Now, we cite some errors that can hinder the tracking process, we will propose for each type a solution to reduce its incidence as much as possible. We have four potential errors cases to appear, that can negatively affect tracking quality.

Case (1): The number of objects in the frame (i-1) is greater than the number of objects in the frame (i). Case (2): The number of objects in the frame (i-1), is less than the number of objects in the frame (i). Case (3): The number of objects in the frame (i-1) is equal to the number of objects in the frame (i). Case (4): Detection of a target object "Sheep" as a non-target object like "Cow". It is often caused by occlusion or detection of the object more than once.

We proposed solutions for almost all detection errors, that give a better annotation video (by detecting objects, identifying target objects, then tracking them through the video). The steps for phase 1, then, are as follows:

1-Detection using Mask R-CNN (Save each image, Extract classification, center coordinates (x, y) and mask matrix of each object). 2-Eliminate double masks: (Compare all mask arrays for single image, Eliminate duplicated object). 3-Propose associations and verify them (Propose an association using the Hungarian algorithm [7], Compare the matrices of each proposed pair, Give each verified pair the same name, Save the pairs have the same name, Search for each odd object, a partner among all previously recorded objects, Give). 4-Eliminate non-target objects: (Compare the classification of each object in all images, Eliminate objects which most of the time were classified as objects other than sheep). 5-Eliminate objects that never move: (Compare the coordinates of each object in all frames, Eliminate objects that never move throughout the

video). 6-Find undetected objects: (Extract the image where the object was not detected and the nearest image from which the object was detected, Predict possible coordinates for the new center, Find the coordinates the closer based on the color of each 7 pixel around the center, save the coordinates if it exists). 7-Empty the centers of same mask:(Count the centers belonging to each mask, in each frame, Change the (x, y) of all the centers of each mask has more than one center, to (0,0)). 8-Delete empty tables.

3.2 Study Knowledge

Ontology plays a central role in our methodology. The ontology serves as a pivot to combine the classification of results and the formalization of knowledge. With a set of individual class instances, the ontology can constitute a knowledge base [11]. In our study, we will express in Web Ontology Language (OWL) the information that helps to specify the behaviors and analyze them, while the semantic rules are expressed in Semantic Web Rule (SWRL).

Knowledge Construction: Specify Animal Behavior. We mean, by the behavior of animals the activities that they perform in the pasture: movement, ingestion, or rest. Given the importance of the results of these activities, we will create an ontology model (Fig. 6) that will help us achieve the best and most accurate results.

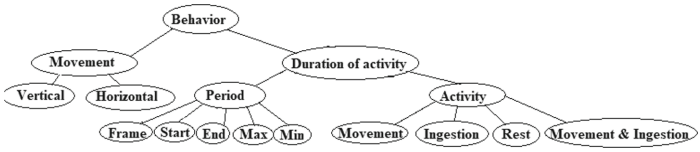

Fig. 6. An extract from animal behavior ontology model.

The semantic rule modeling process includes building Mark rules and decision rules. The construction of Mark rules is based on a semantic concept, and the process moves from low-level features to semantic concepts. Then, decision rules are obtained based on Mark rules and prior knowledge. The process moves from advanced features to identifying trends, period features (frame, start, end, minimum, maximum), and activities. The ontology model of Mark rules represents the different states of motion on the X and Y axes, with each subclass representing a state of the parent class.

Mark rules are expressed in SWRL and semantic relationships between object features and classes are constructed. For example, the case where $Y_D > Y_F$ and $X_D > X_F$ is expressed in SWRL as follows:

- periode(?I,?Y_D,?Y_F,?Y_{min},?Y_{max}), greaterThan(?Y_D,?Y_F)→ $Y_D > Y_F$
- periode(?I,?X_D,?X_F,?X_{min},?X_{max}), greaterThan(?X_D,?X_F)→ $X_D > X_F$

This means that the characteristic Y_D of an object > their Y_F designates the case we called "$Y_D > Y_F$", the same for the characteristic X_D of an object > their X_F denotes the case we called "$X_D > X_F$". With, period(? X), X is an individual of period, greaterThan(?X, ?Y) represents attributes, and x and y are variables.

The decision rules for four types of activities are acquired from a priori knowledge and the technical regulations of the project. We have formalized these decision rules using OWL as follows:

- Movement $\equiv X_D > X_F \sqcup X_D < X_F$
- Ingestion $\equiv Y_D > Y_F \sqcup Y_{min} > Y_D \sqcup Y_i > Y_{max}$
- Rest $\equiv \neg$ Movement $\sqcap \neg$ Ingestion
- Movement&Ingestion $\equiv$ Movement $\sqcap$ Ingestion

Decision rules are expressed in SWRL, and semantic relationships between mark rules and classes are constructed. For example, Movement&Ingestion is expressed in SWRL as follows:

$X_D > X_F$(?period), $Y_D > Y_F$(?period) $\rightarrow$ Displacement&Ingestion(?period). This means that an activity with characteristics (movement on the X axis and movement on the Y axis) $X_D > X_F$ and $Y_D > Y_F$ is Movement&Ingest.

Knowledge Analysis: Study Animal Behavior. The objectives of our study of animal behavior in pastures are to know (1) the effect of grazing on natural rangelands and (2) the role of rational grazing in animal self-sufficiency. Using the results of the "Specify Behaviors" process and set of rules we can build an ontology that allows us to achieve our goals. We explain here one of these extraction knowledge processes.

Use of Pasture: Three main criteria to know the state of pastures degradation, whether it is very poor, poor, moderately rich, rich, or satisfactory. The ontology mark rules model presents the TRM rate of restoration of the natural vegetation cycle, the season, and the animal behavior (mainly represented by the consumption rate of pastoral species; ingestion rate).

Mark rules are expressed in SWRL and semantic relationships between object features and classes are constructed. For example, the case where the TRM rate is class [0.20], the season is "summer" and the ingestion rate is]40.50] are expressed in SWRL as follows:

- TRM(?rate), greaterThan(?rate, 0), lessThanOrEqual(?rate, 20) $\rightarrow$ [0,20]
- Tingestion(?$rate_{ing}$), greaterThan(?$rate_{ing}$, 40), lessThanOrEqual (?$rate_{ing}$, 50) $\rightarrow$]40,50]
- Season(?S), Equal (?S, summer) $\rightarrow$ summer

This means that the "rate" characteristic of an object > at 0 and < at 20 denotes the case we called "[0,20]", and the characteristic "$rate_{ing}$ " of an object > at 40 and < at 50 designates the case that we called "]40,50]", the same for the characteristic "S" of an object = summer designates the case that we called "summer" With, P(? X), X is an individual of period, lessThanOrEqual(?x, ?y) represents attributes, and x and y are variables. Decision rules for pasture use states

are acquired from a priori knowledge and technical regulations of the project. We formalized the decision rule from the previous example, using OWL as follows:

- Massive ≡ summer ⊓ [0.20] ⊓]40.50]

Decision rules are expressed in SWRL, and semantic relationships between brand rules and classes are constructed. For example, Rich usage is expressed in SWRL as follows: summer(?S), [0.20] (?rate),]40.50] ($?rate_{ing}$) → Massive(?state).

This means that a use state with summer characteristics, [0.20](TRM) and]40.50] (ingestion rate) is Heavy use.

4 Evaluation and Discussion

Our approach has resolved many issues, both in detection and tracking. Here we present the challenges we discussed earlier and the results obtained.

Detection. Our approach manages to detect and recognize all target objects well. It overcame the challenges caused by occlusion, and the variance of objects number throughout the video.

Association. Since our approach has overcome the challenge of varying the number of objects in the video, it can track every object well. In addition, the deleting of object coordinates when it has a large percentage of mask intersection with another helps to avoid calculating the object more than once in each frame and thus it provides good and accurate tracking

Tracking. To test the effectiveness of our approach, we used two videos, one (video 1) of 28 s, contains 5 sheep present since the beginning of the video sequence, and another sheep has been entered at the end of the video (Fig. 7(a). The second (video 2) is 5 s long, and contains more than 25 sheep present from the beginning of the video sequence to the end, as shown in the Fig. 7(b). The first video was converted to 571 frames, Fig. 8V(1).(a) shows the results of detection and association steps. The number of sheep detected is greater than the actual number of sheep. Figure 8V(1).(b) shows our proposed approach results to find the missing coordinates while deleting the repeated ones, so the alleged objects were deleted. It contained the coordinates of several sheep, and this happened because of occlusion between the sheep. Our approach here was able to track all the six sheep by 100%.

Fig. 7. One of video_1 (a) and video_2 (b) frames.

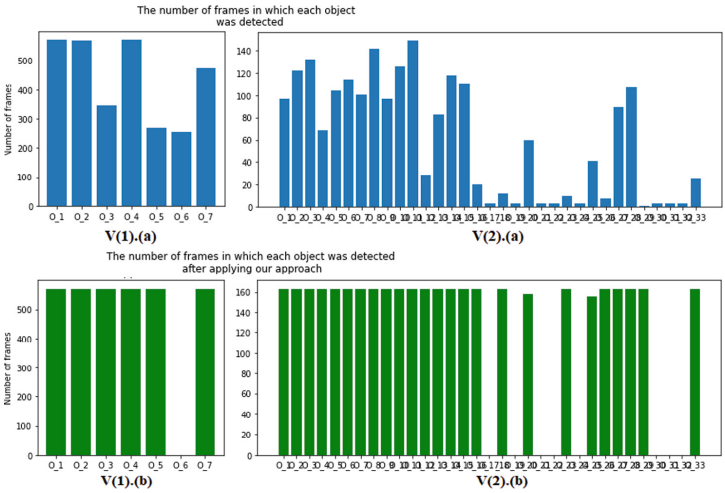

Fig. 8. Object detection results using Mask RCNN.

The video (2) was converted to 163 frames, the detection and the association result was shown in Fig. 8V(2).(a). 33 sheep were found, this number is higher than the actual number of sheep. The sheep that did not move their coordinates throughout the detection period were removed, Fig. 8V(2).(b) shows the result of our optimization.

We noticed in this case (video 2), that there was great confusion in the association between certain objects because they confused their coordinates with the other's coordinates. This happened when there was a large occlusion in addition to the rapid movement of these sheep.

5 Conclusion

In this article, we introduced our two-stages approach. The first is the semantic annotations of a video sequence to extract all the information needed to build and study the knowledge in the second stage. We clarified all the steps that guarantee us a 100% tracking based on a 100% detection using the Mask RCNN detector and a true association by pressing the Hungarian algorithm to reduce the process of looking for associations. In each step, we mentioned the problems that could be hindering our work and the appropriate solutions to remove them as much as possible.

Although object detection had a lot of attention in the past decade, the best detectors are still far from saturated in performance, it also remains an open challenge. In this paper, we presented a new method for video object tracking based on the Mask R-CNN detector. The result of this approach helps us to assure a good annotation of the video. Based on it, in the second stage, we will study the behavior of each animal and determine the effects of the behavior on the animal itself, and the pasture.

References

1. Ciaparrone, G., et al.: Deep learning in video multi-object tracking: a survey. Neurocomputing **381**, 61–88 (2020)
2. Girshick, R., et al.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 580–587 (2014)
3. Gu, H., et al.: An object-based semantic classification method for high resolution remote sensing imagery using ontology. Remote Sens. **9**(4), 329 (2017)
4. Guerrero, A., Villagrá, V.A., de Vergara, J.E.L., Berrocal, J.: Ontology-based integration of management behaviour and information definitions using SWRL and OWL. In: Schönwälder, J., Serrat, J. (eds.) DSOM 2005. LNCS, vol. 3775, pp. 12–23. Springer, Heidelberg (2005). https://doi.org/10.1007/11568285_2
5. Jiao, L., et al.: New generation deep learning for video object detection: a survey. IEEE Trans. Neural Netw. Learn. Syst. (2021)
6. Kim, S.J., et al.: Online tracker optimization for multi-pedestrian tracking using a moving vehicle camera. IEEE Access **6**, 48675–48687 (2018)
7. Kuhn, H.W.: The Hungarian method for the assignment problem. Nav. Res. Logist. Q. **2**(1–2), 83–97 (1955)
8. Lin, T.Y., et al.: Feature pyramid networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2117–2125 (2017)
9. Liu, L., et al.: Deep learning for generic object detection: a survey. Int. J. Comput. Vision **128**(2), 261–318 (2020)
10. Luo, W., et al.: Multiple object tracking: a literature review. Artif. Intell. **293**, 103448 (2021)
11. Noy, N.F., et al.: Ontology development 101: a guide to creating your first ontology (2001)
12. Parekh, H.S., et al.: A survey on object detection and tracking methods. Int. J. Innov. Res. Comput. Commun. Eng. **2**(2), 2970–2979 (2014)
13. Phan, N., et al.: Ontology-based deep learning for human behavior prediction with explanations in health social networks. Inf. Sci. **384**, 298–313 (2017)
14. Sasse, J., et al.: Semantic metadata annotation services in the biomedical domain-a literature review. Appl. Sci. **12**(2), 796 (2022)
15. Wang, X., et al.: Data modeling and evaluation of deep semantic annotation for cultural heritage images. J. Doc. (2021)
16. Welty, C., Guarino, N.: Supporting ontological analysis of taxonomic relationships. Data Know. Eng. **39**(1), 51–74 (2001)
17. Wen, Y., et al.: Semantic modelling of ship behavior in harbor based on ontology and dynamic Bayesian network. ISPRS Int. J. Geo Inf. **8**(3), 107 (2019)
18. Yao, R., et al.: Video object segmentation and tracking: a survey. arXiv preprint arXiv:1904.09172 (2019)
19. Yazdi, M., Bouwmans, T.: New trends on moving object detection in video images captured by a moving camera: a survey. Comput. Sci. Rev. **28**, 157–177 (2018)
20. Zaidi, S.S.A., et al.: A survey of modern deep learning based object detection models. Digit. Signal. Process. 103514 (2022)
21. Zhou, Z., et al.: Online multi-target tracking with tensor-based high-order graph matching. In: 2018 24th International Conference on Pattern Recognition (ICPR), pp. 1809–1814. IEEE (2018)