Citation: Sara Guillén-Garde, Gonzalo López-Nicolás, and Rosario Aragüés. Detection and tracking of livestock herds from aerial video sequences. In Danilo Tardioli, Vicente Matellán, Guillermo Heredia, Manuel F. Silva, and Lino Marques, editors, ROBOT2022: Fifth Iberian Robotics Conference, pages 423–434, Cham, 2023. Springer International Publishing.

Detection and tracking of livestock herds from aerial video sequences

Sara Guillén-Garde, Gonzalo López-Nicolás, Rosario Aragüés

Instituto de Investigación en Ingeniería de Aragón, Universidad de Zaragoza, Spain
raragues@unizar.es and gonlopez@unizar.es

Abstract. Autonomous herding research is becoming increasingly relevant. In this work, a model for sheep detection in herds from aerial video sequences is proposed, using the convolutional neural network Mask R-CNN. Several trainings with different datasets have been performed for achieving the model. An improvement in the detection metrics, through a visual tracking tool, allows not only detecting the individual sheeps in the herd, but also tracking them along the different frames in aerial video sequences. This system could be used, for example, in a drone to carry out livestock supervision, in addition to obtaining metrics that allow knowing the status of the herd. Finally, the method has been validated using several tests on images and videos of livestock in real outdoor environments.

1 Introduction

This work is framed within the context of research on methods of supervision, monitoring and automation of agricultural and livestock tasks. Specifically, livestock monitoring and autonomous herding scenarios are considered. While the demand for products, and the requirements of animal welfare are increasing, there is a shortage and aging of workers in the livestock industry, due to the depopulation of rural areas. In this context, the research and development of new technological solutions for the automation of livestock tasks become necessary. These initiatives will provide better working conditions for ranchers and farmers, and will make livestock activities more efficient.

Some tasks whose automation can facilitate the work of shepherds and that are already being investigated, are for example: the automated classification of different types of cattle [19] using a drone camera, or the counting of cattle on a farm or herd in real time [17]. The automation of herding is also being investigated. The possibility of using an unmanned drone as a sheep herder has been studied in [20], concluding that this is a realistic option, since it was shown that sheeps react appropriately to the drone, without getting more stressed than when using dogs. Using multi-robot teams to observe the herd and to cooperatively guide it is also of interest to the community [9] [14].

In order to address these tasks, it is critical to count on highly accurate perception mechanisms that allow locating and tracking the position of the animals and extracting useful information from the herd state. Machine learning tools, specifically convolutional neural networks, are very useful in this area. They can be used for livestock detection on any device with a camera, as they allow object detection in images and videos. Note that the animal detection and identification problem presents several challenges, such as the overlaps between animals, the obstacles such as rocks and trees that create occlusions, or the similarity of color between the animals and the background. The rapid development of object detection in deep learning provides techniques with promising results [4] [12] [11]. Several studies on animal detection, segmentation and counting using convolutional neural network detectors have come out [19] [1] [10] [18]. In this work, machine learning will be used to detect sheeps in aerial perspective images and videos. Specifically, we use the convolutional neural network Mask R-CNN [4].

A key point of the proposed work is the combination of these machine learning techniques, with computer vision and visual tracking tools. An existing multi-object tracking tool will be adapted to use our detector, thereby tracking the sheeps in a herd. In addition, the result of the tracking and other visual information of the image will be used to obtain metrics of the herd that allow evaluating its status. Note that livestock herds follow collective behaviors that are classical in the multi-agent literature, such as the well known Boids model [13]. Here, in particular, we obtain metrics of the herd state related with the Voronoi tessellation, that often appears also in the multi-robot literature [16]. Livestock tracking has already been studied previously using different techniques, such as thermal sensors, GPS systems and cameras [5] [15] [7].

This paper is organized as follows. Section 2 details the datasets used, the training process and the evaluation of the detection models. Section 3 presents the procedure for detection and visual tracking in videos. Some examples of potential applications for the system are shown in section 4. Finally, section 5 presents the conclusions. A video summary of all the work carried out, which also shows the operation of the system in aerial videos is provided [1].

2 Datasets and training

2.1 Model evolution

This section will briefly summarize the models obtained before reaching the final one, and the evaluation of the final model will be shown. The convolutional neural network we used for sheep detection is Mask R-CNN [4], since it is one of the best and most used methods. To take advantage of transfer learning, a pre-trained model with the entire COCO Dataset [6] has been used as the basis for training. The dataset contains 91 different classes and 200K images. For the evaluation of the models obtained with the different trainings, the following four metrics are used. *Precision* represents the percentage of all detections which are actually sheeps. *Recall* represents the percentage of all the sheep in the images that our model has detected. The *Precision-Recall curve* is a graph that shows the tradeoff between precision and recall for different probability thresholds. The

[1] Detection and tracking of livestock herds. https://youtu.be/5DtY76MA2OI

Average Precision (AP) represents the area under the Precision-Recall curve. The larger the area under the curve, the higher precision and recall values can be obtained at the same time.

The dataset used in the training of the first model was the COCO dataset. The 1.6k images with sheep class segmentation labels it contains were used in training. The model resulting from training with these images is not capable of correctly detecting sheeps in aerial perspective. This is because the images from the COCO dataset are mostly photos taken near the sheeps and from ground perspectives (see left images in Fig. 1). Object detectors trained with conventional images do not work well on aerial images. Note that the appearance of all kinds of objects, and specifically of sheeps, is very different in both perspectives, as shown in Fig. 1. In Fig. 3, the detections made on the same image by the different trained models are shown. The first image corresponds to the detections of this first model. It can be seen that, although the model detects several sheeps, it is not able to identify their area correctly, and there are still many false negatives. The average precision, AP, value of this model was 44.26%.

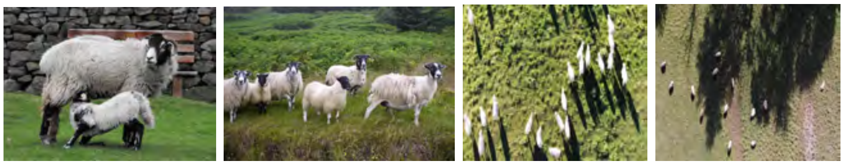


Fig. 1: Images on the left side are examples from the COCO Dataset of sheeps in ground perspective. Images on the right side are examples of sheeps in aerial perspective from the Aerial Sheep Dataset.

In order to improve the previous model, and get it to learn the characteristics of sheeps seen in aerial perspective, all the images of the only dataset found of sheeps in aerial perspective, the Aerial Sheep Dataset [2], with 4.1k annotated images, were added to the training. However, this dataset has several drawbacks: the resolution and quality of the images are low, in some images sheeps are so far away and close together that they are almost indistinguishable, and the quality of the labels is low and irregular. Two examples of images from the dataset with low resolution, indistinguishable sheeps, and low quality labels can be seen in Fig. 2. These problems will limit the quality of the model. The result of this training, unlike the previous one, does manage to correctly detect sheeps in aerial images. For this reason, the AP value of this model increased to 80.30%. The second image in Fig. 3 shows examples of detections made with this model.

Despite being able to correctly detect sheeps on aerial view images, there are still false negatives and positives. The false positives are probably due to two

[2] Aerial Sheep Image Dataset.https://universe.roboflow.com/riis/aerial-sheep

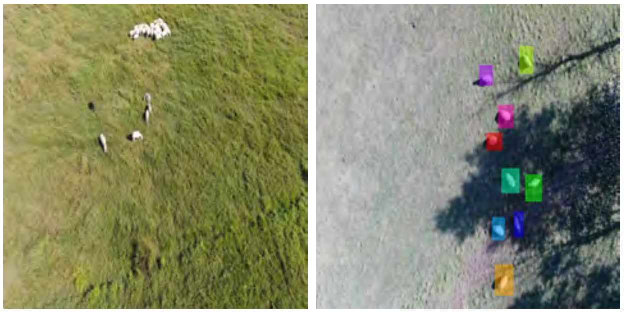


Fig. 2: Examples of images from the aerial dataset where sheeps are barely visible (left), or with poor quality labels that are too loose (right).

factors, the first one being the quality of the training images and their labels. And the second, the absence of true negative examples in the training images, since most of the images contain only sheep and meadow. So, to reduce the number of false positives, images with true negatives were added to the training images [8]. COCO images with the classes dogs, cows and people, were added to the training, a total of 800 images. In this way, the neural network would learn which objects are not sheeps. The improvement in this model was small, but false positives have been reduced, and therefore the AP has increased to 81.56%

Despite the improvements achieved, it is considered that this model is limited by the quality of the main dataset, the Aerial Sheep Dataset. So, the 550 best quality images of the Aerial Sheep Dataset were chosen, in which sheeps could be better distinguished. In addition, 40 unlabeled images obtained from video frames were added. Those images had more types of sheeps and higher resolution. Finally, all images were manually re-labeled, to obtain well-fitting labels for the sheeps. The resulting model obtained the best results, despite having used fewer images for training than with the previous models explained in this section (about 10% of images). The images of the dataset have been divided 80% for training and 20% for testing. And within the training ones, a 20% has been reserved for validation. Regarding the details of the training, it has been carried out in two stages: first, a training of the head layers for 10 epochs, with a learning rate of 0.001. In this stage, the training was performed only on the layers that had been initialized randomly, that is, the layers that were not used from the COCO pre-trained model. Then, a complete network training for 5 epochs, with a learning rate of 0.0001. These parameters were fitted experimentally, seeking to minimize detection errors in both validation and test data.

For the evaluation of this model, the threshold has been adjusted experimentally, concluding that the best value is 0.7. With the chosen threshold, the precision reaches a value of 93.82%, and the recall of 90.25%. There is a reduction in false positives and negatives, which is reflected in a precision, recall, and AP increase, as seen in Table 1, compared to the AP values of the previous model.

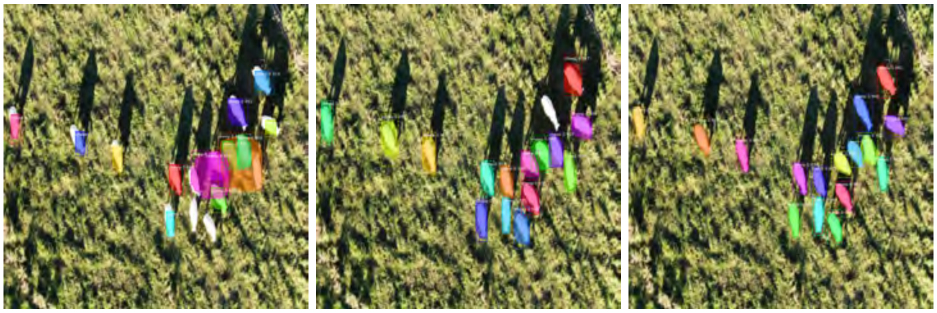


Fig. 3: Comparison of the detections made with three different models in the same image. It can be seen that each new training obtains better detections (from left to right).

This model is the definitive one, and the one that will be used for tracking. The third image in Fig. 3 shows the detections made by this model on the same image as in previous models. The comparison between the detections of the three models shows that the last one reduces the false negatives and positives, and obtains well-fitting segmentation masks for detections.

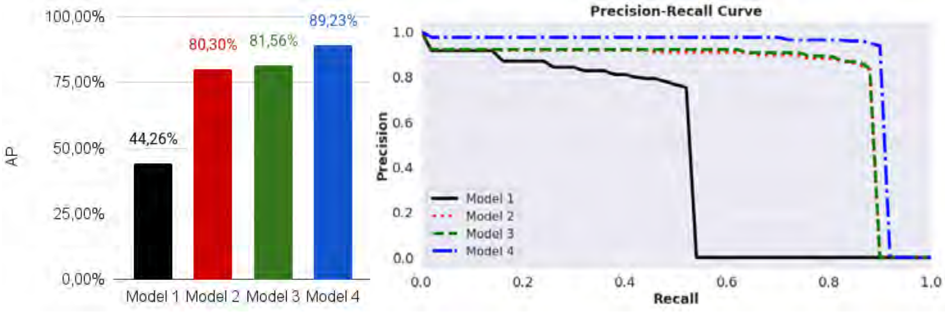
Table 1: Table of the final model metrics in sheep detection, obtained with a confidence threshold of 70%. It also includes the model AP value.

Metric	Value
Precision	0.9382
Recall	0.9025
AP	0.8923

Fig. 4 shows a comparison of the AP values of the four models explained, and its Precision-Recall curve. Models 1, 2 and 3 correspond to those discussed at the beginning of this section, and model 4 is the final model. These graphs allow us to see how, with each new training, the area under the Precision-Recall curve increases, allowing higher precision and recall values to be reached at the same time, and producing an increase in the AP value.

3 Visual tracking

Once good results were obtained in the detection in images, the proposed model was extended to process sequences of images from videos, since the first step



(a) (b)

Fig. 4: (a) Comparative graph of the AP values of the four trained models. (b) Comparative graph of the Precision-Recall curves of the four models.

to carry out the tracking is the detection in videos. All videos used have been downloaded from Envato Elements [3]. To evaluate the performance of the detection model, several videos were labeled manually. Fig. 5 shows a frame of a video on which the detection has been performed. On the left side is the video with the model detections, which are green if they are correct, and red if they are false positives. On the right is the video with the ground-truth labels. In the image we see that most of sheeps have been correctly detected. There is only one false positive, placed in the left part of the image, in which three sheeps have been wrongly mixed up. And there is a false negative (a sheep that has not been detected) in the center-bottom part of the image.

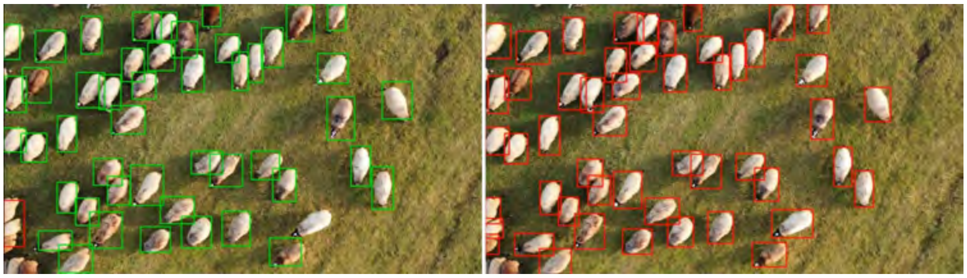


Fig. 5: The images on the left show the detections made by our model, and the ones on the right show the ground-truth labels.

The next step in the proposal is to use a visual tracking tool to be able to identify each sheep in the herd, and to track each one, recognizing it in each frame of the video. The tool used was SORT tracker [2] (Simple Online Real-

[3] Envato Elements. https://elements.envato.com/es/

Time Tracking). It was chosen experimentally among other tracking techniques widely used, several were tested in the context of videos used in the project, and SORT obtained the best performance. SORT is a 2D tracking algorithm for multiple objects in video sequences. The high level operation of the tracker would be as follows: the Kalman filter is used to predict the location of objects in a video frame. Then, the Hungarian algorithm assigns these predictions to the detected objects, assigning to each prediction the detection that maximizes the Intersection Over Union metric (IoU). This IoU metric measures the amount of overlapping of the bounding boxes of the predicted and detected objects. When a detection is associated with one of the objects, the state of the object is updated with the position of the detection's bounding box, and based on the updated state, the Kalman filter predicts the speed, position, and geometry of the bounding box in the next frame. If no detection is associated with the target, its state is simply predicted based on the uncorrected state.

The videos recorded with a drone usually suffer from changes in perspective, changes in height, and camera shakes and flickers. In order to be able to perform the sheep visual tracking, it is convenient to have minor changes of perspective as well as having a perpendicular perspective to the ground. Also, for being able to approximate distances in pixels to real distances, the perpendicular perspective is useful. For this reason, a program that allows to rectify the perspective in videos has been developed, so that the perspective is as similar as possible to the one described. A widely used rectification technique based on homography has been used, a detailed explanation of the technique can be found at [3]. Two frames, each one from a perspective-corrected video, are shown in Fig. 6. The original frame is shown on the left side and corrected on the right side. Note that the right part (corrected images) looks like if the camera was normal to the surface. An example can be seen in the project video[1].

Fig. 7 shows four frames, all of them from different moments of a sheep tracking example video. Each ID represents a tracked sheep, and the color of the bounding boxes is associated with each ID, so it remains constant. Those frames show that most of the sheeps are correctly tracked. An example would be sheeps number 27 and 41, which are marked with a circle. Sheep number 40, which is marked with a cross, is an example of tracking failure, since the tracker loses it in the third image, and then it finds the sheep again in the fourth image, but wrongly assigns another ID to it. To see the sheep tracking on video and to appreciate its performance, watch the summary video[1].

3.1 Evaluation of detection and tracking in videos

For the evaluation of the detection in videos, the average precision and recall values for the video have been measured, using the bounding boxes returned by the CNN, not the segmentation, so that the comparison with the tracking metrics is fair. They appear in the second column of the Table 2. The threshold used to calculate them has been the same as in the evaluation of the model, i.e., 0.7. The precision value obtained was almost 99%, without the recall falling below 92-93%. Next, to evaluate the sheep tracking performance, four metrics

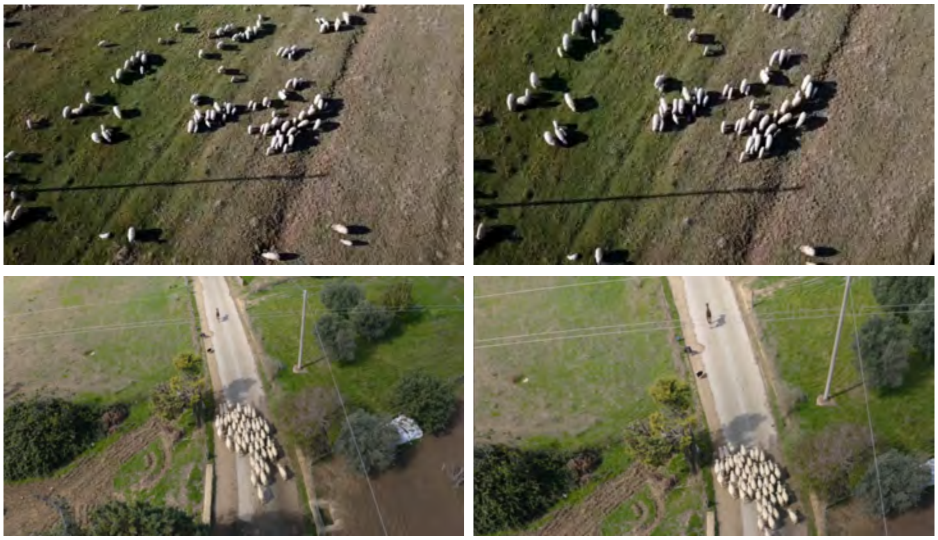


Fig. 6: Left: Original video frames. Right: Perspective rectification of the video frames.

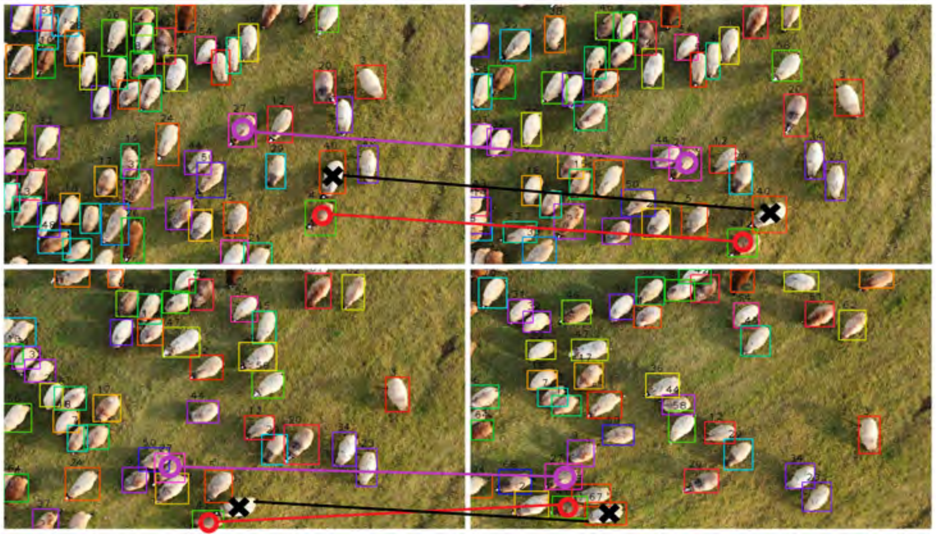


Fig. 7: Four frames of different moments of a sheep tracking example video. Each ID represents a tracked sheep, and the color of the bounding boxes is associated with each ID, so it remains constant.

have been calculated: the detection precision and recall, and the association precision and recall. The tracker detection metrics are calculated the same as the model detection ones, but instead of using the bounding boxes from the detection model, the bounding boxes predicted by the tracker are used. They are shown in the third column of the Table 2. The association metrics measure whether the tracker is capable of associating all detections of the same sheep, to the same tracked object. So, they evaluate if whenever the sheep appears in a frame, the tracker identifies that sheep and does not confuse it with another one. The association metrics appear in the fourth column of the Table 2.

Comparing the precision and recall values obtained for detection without tracking, detection with tracking and association, we can see that the detection recall value has improved thanks to tracking. The reason is that in frames where the detector is not able to detect a sheep, including cases of occlusion, the tracker often correctly predicts the bounding box of the sheep. In addition, the increase in recall does not produce a drop in precision, which remains at a value of 98%. Furthermore, as this video has a stable and perpendicular perspective, and all sheeps can be distinguished, the tracker's association measurements are also high, with a precision of 93,4% a recall of 90%.

Table 2: Mean values of precision, recall of detection without tracking, detection with tracking and association.

Metric	Detection without tracking	Detection with tracking	Association
Precision	0.9891	0.9831	0.9340
Recall	0.9275	0.9500	0.9024

4 Potential applications for the system

The developed system opens up possibilities for the future to facilitate autonomous herding. In this section, we show some examples of metrics that can be extracted from Voronoi Diagrams of the animal locations. Future extensions of this work will use these metrics to extract relevant information of the state of the herd and use this information for a proper monitoring and herding of the livestock. Note that in order to obtain the metrics and Voronoi Diagrams, it is required to detect and locate the individual animals, as we explained in the previous sections. Note also that, in videos where perspective is perpendicular to the floor of the video, distances in pixels can be easily approximated to distances in meters, using an approximate measurement of the area of a sheep, 1×0.33 m^2. The evolution of the Voronoi diagram has been obtained for several videos, and in these diagrams, the Voronoi regions have been colored according to their area. Regions with an area less than 1 m^2 are colored red, those with an area between 1 and 20 m^2 are blue, and those with an area greater than 20 m^2 are green. Fig. 8 shows a frame from an example video, with the Voronoi diagram for that frame superimposed.

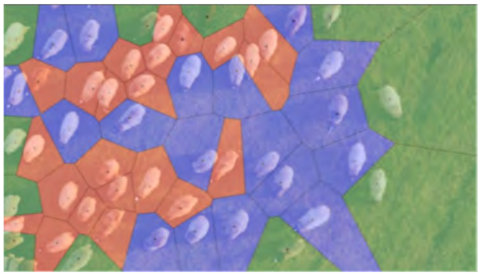


Fig. 8: Example of a Voronoi diagram calculated on a video.

From the Voronoi diagram and other values measured in each frame, the following metrics can be calculated during a video:

- **Area of the herd (A):** Area in m^2 that the visible part of the herd occupies.
- **Density (D):** Number of sheep per square meter, inside the herd.
- **Mean distance to centroid (DC):** Average distance in meters to the herd centroid.
- **Farthest sheep (FS):** Maximum distance to the herd centroid, which corresponds to the farthest sheep from the herd, which is also the animal with the highest risk of splitting up from the herd and eventually getting lost.
- **Mean distance between sheep (DS):** Average distance between sheeps in meters.
- **Space Occupancy (SO):** Percentage of space occupied by sheeps, inside the herd.
- **Area of the Voronoi regions (AV):** The mean area of the regions of the Voronoi diagram has been measured in m^2, inside the herd.
- **Ratio of maximum area to minimum area (RA):** The ratio divides the maximum area that a sheep has in the herd, against the minimum area. This value guides how homogeneous the distribution of space in the herd is.

Table 3 shows the average values of those metrics for the video shown in [1], and Fig. 9 shows a graph of the evolution of the metrics during the video. With these examples we can see that relevant information about herds can be obtained through detection and tracking. This information can be interesting for making decisions in autonomous herding, for knowing, for example, which sheep is further away, and if it is too far, redirect it to the herd. Or, if the distance between sheep is too small, or the density is too high, move the flock to a larger place. If, on the contrary, the distance is too big, it is necessary to bring the sheep closer together, so that they do not disperse.

Table 3: Average metrics obtained for the video.

A (m^2)	D (sheeps/m^2)	DC (m)	FS (m)	DS (m)	SO (%)	AV (m^2)	RA
39.74	1.2164	2.7403	5.35	3.7784	0.4054	1.9372	0.0327

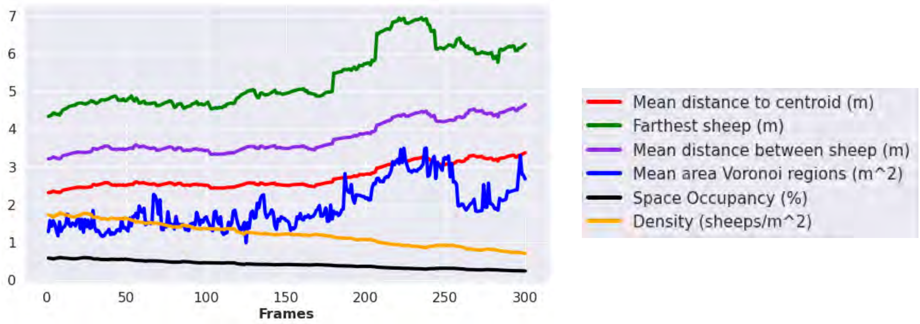


Fig. 9: Metric values measured from the video.

5 Conclusion

In this work, several automated models for sheep detection and segmentation in aerial images have been carried out and, after several tests with different datasets, a model with an average precision of almost 90% has been developed. Next, the model has been adapted to perform the detection on aerial video sequences, obtaining a fairly good mean precision value, although a slightly lower recall value. Subsequently, it has been possible to improve the detection of the model, specifically its recall, using sheep tracking. The model resulting from this combination of techniques allows tracking the sheeps of the herd, and obtaining metrics that report the status of the herd. The proposed system could replace the manual observation of cattle, and facilitate the work of the shepherd. Future work could integrate the proposed algorithm in a drone, and form a multi-robot system with several drones to supervise sheeps and dog-like robots to herd sheeps. Another line of study could be to improve the current model using a larger and more varied dataset and compare its performance with other different deep learning techniques such as attendance models and transformers.

Acknowledgment

This work was supported by the Spanish Government/European Union through projects PGC2018-098719-B-I00 and PID2021-124137OB-I00 funded by MCIN/AEI/10.13039/501100011033.

References

1. Ardö, H., Guzhva, O., Nilsson, M., Herlin, A.H.: Convolutional neural network-based cow interaction watchdog. IET Computer Vision **12**(2), 171–177 (2018)
2. Bewley, A., Ge, Z., Ott, L., Ramos, F., Upcroft, B.: Simple online and realtime tracking. In: IEEE Int. Conf. Image Processing, pp. 3464–3468 (2016)
3. Hartley, R., Zisserman, A.: Multiple View Geometry in Computer Vision, 2 edn. Cambridge University Press (2004)
4. He, K., Gkioxari, G., Dollar, P., Girshick, R.: Mask R-CNN. In: IEEE Int. Conf. Computer Vision, pp. 2961–2969 (2017)
5. Kim, W., Cho, Y.B., Lee, S.: Thermal sensor-based multiple object tracking for intelligent livestock breeding. IEEE Access **5**, 27,453–27,463 (2017)
6. Lin, T., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common objects in context. In: European Conf. Computer Vision, pp. 740–755. Springer (2014)
7. Molapo, N.A., Malekian, R., Nair, L.: Real-time livestock tracking system with integration of sensors and beacon navigation. Wireless Personal Communications **104**(2), 853–879 (2019)
8. Ng, A.: Machine Learning Yearning (2018)
9. Pierson, A., Schwager, M.: Bio-inspired non-cooperative multi-robot herding. In: IEEE Int. Conf. Robotics and Automation, pp. 1843–1849 (2015)
10. Qiao, Y., Truman, M., Sukkarieh, S.: Cattle segmentation and contour extraction based on Mask R-CNN for precision livestock farming. Computers and Electronics in Agriculture **165**, 104,958 (2019)
11. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: IEEE Conf. Computer Vision and Pattern Recognition, pp. 779–788 (2016)
12. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence **39**(6), 1137–1149 (2017)
13. Reynolds, C.: Flocks, herds and schools: A distributed behavioral model. Conf. Computer Graphics and Interactive Techniques. Association for Computing Machinery pp. 25–34 (1987)
14. Sebastián, E., Montijano, E.: Multi-robot implicit control of herds. In: IEEE Int. Conf. Robotics and Automation, pp. 1601–1607 (2021)
15. Sergeant, D., Boyle, R., Forbes, M.: Computer visual tracking of poultry. Computers and Electronics in Agriculture **21**(1), 1–18 (1998)
16. Teruel E., A.R.L.N.G.: A practical method to cover evenly a dynamic region with a swarm. IEEE Robotics and Automation Letters **6**(2), 1359–1366 (2021)
17. Tian, M., Guo, H., Chen, H., Wang, Q., Long, C., Ma, Y.: Automated pig counting using deep learning. Computers and Electronics in Agriculture **163**, 104,840 (2019)
18. Xu, B., Wang, W., Falzon, G., Kwan, P., Guo, L., Chen, G., Tait, A., Schneider, D.: Automated cattle counting using Mask R-CNN in quadcopter vision system. Computers and Electronics in Agriculture **171**, 105,300 (2020)
19. Xu, B., Wang, W., Falzon, G., Kwan, P., Guo, L., Sun, Z., Li, C.: Livestock classification and counting in quadcopter aerial images using Mask R-CNN. International Journal of Remote Sensing **41**(21), 8121–8142 (2020)
20. Yaxley, K.J., Joiner, K.F., Abbass, H.: Drone approach parameters leading to lower stress sheep flocking and movement: sky shepherding. Scientific reports **11**(1), 1–9 (2021)