DEVELOPING LIVESTOCK DETECTION MODEL FOR VERY HIGH-RESOLUTION SATELLITE IMAGERY USING AERIAL IMAGERY AND DEEP LEARNING IN KENYA

Ian Ocholla [a,b], Petri Pellikka [a,b], Faith Karanja [c], Mark Boitt [d], Tuomas Väisänen [a], Ilja Vuorinne [a,b], Janne Heiskanen [a,e]*

[a] Department of Geosciences and Geography, University of Helsinki, Finland

[b] Institute for Atmospheric and Earth System Research, University of Helsinki, Finland

[c] Department of Geospatial and Space Technology, University of Nairobi, Kenya

[d] Institute of Geomatics, GIS & Remote Sensing, Dedan Kimathi University of Technology, Kenya

[e] Finnish Meteorological Institute, Helsinki, Finland

*ian.ocholla@helsinki.fi

ABSTRACT

Space-based livestock counting coupled with big data analytics, provides novel ways for rangeland management and monitoring of greenhouse gas emissions. However, livestock are difficult to distinguish from space. We test the feasibility of YOLOv5 algorithm to detect cattle from satellite imagery based on a model trained on 10cm aerial imagery. Four YOLOv5 architectures were assessed on four image datasets. YOLOv5x model with inclusion of augmented and background data had the highest accuracy of 51.8% and the lowest count error (CE) of -5.6% relative to manual annotation. Then, the model was tested on a downsampled dataset simulating very high-resolution (VHR) satellite imagery at 30cm resolution. YOLOv5x accuracy declined by 5.5% and CE underestimation increased to -28.9%. The results show that a simple one-stage detector trained on higher resolution images can be an effective approach to detect cattle from satellite imagery although further studies on real VHR satellite imagery are needed.

***Index Terms*—** VHR aerial, livestock, object detection, YOLOv5

1. INTRODUCTION

Accurate livestock counts are vital for sustainable grazing management in rangelands. Approximately 50% of earth land mass is occupied by rangelands, which host a billion of the global livestock and several hundreds of million pastoralist households [1]. Currently, these pastoralists are facing imminent challenges in survival of their livestock, from the long drought periods, conversion of pastures to cropland and overstocking [1]. To counter these challenges, proper and up to date livestock counts are essential for pasture management and to curb pastoralists from loss of their livestock during prolonged droughts.

Recent advancements in remote sensing and detection techniques based in computer vision and deep learning have potential to address these challenges. Images for livestock census are generated from unmanned aerial vehicles (UAVs), manned aircrafts, and very high resolution (VHR) satellite sensors. UAVs have transformed collection of remotely sensed data as they can easily maneuver inaccessible areas and are cost friendly; however, they are limited to covering small areas [2]. In contrast, VHR satellites, such as Pleiades NEO, offer spatial resolution of up to 30cm and are a valuable source of big data for detecting livestock regularly over larger areas. However, with such resolution, the livestock still occupies only a few pixels making it difficult to distinguish livestock from the complex background [3]. Therefore, the direct labeling of the training data from VHR imagery is also a challenge.

Popular deep learning techniques for object detection include the one stage detectors; YOLO [4], SSD [5] and Fast RCNN [6]; two stage detectors RCNN [7] and Faster RCNN [8]; and instantaneous segmentation Mask RCNN [9]. Whilst the two stage and Mask RCNN detectors produce better accuracy [10], they often require high computation cost. In contrast, one stage detectors trade speed for accuracy and less complicated [4]. Among the one-stage detectors, YOLO has evolved to have better accuracy compared to SSD and even to two-stage Faster RCNN.

YOLO model has been used in detecting livestock using images captured by low flying UAVs, with high resolution capable to classify multispecies and even distinguish them from the background [11]. However, these studies are confined to research fields, which host limited, and specific livestock breeds and have a homogeneous background [11], [12]. None of the studies have been conducted in rangelands, which are characterized by

doi:10.2760/46796

heterogeneous background, various cattle breeds, and sparsely distributed pastoralists communities. In this study, we compare four different YOLOv5 architectures in detecting and counting livestock in aerial imagery and test the feasibility of upscaling YOLOv5 to detect and count livestock from simulated VHR satellite imagery.

2. MATERIAL AND METHODS

2.1. Study area

The study area is Lumo conservancy (-3° 28' 12" S, 38° 11'31" E), located in Taita Taveta County in the southeastern part of Kenya. The conservancy covers an area of 95km^2 and is characterized by semi-arid to arid environment. It hosts both livestock (cattle, sheep, and goats) from the local community and wildlife.

2.2. Aerial imagery

Aerial survey was conducted in February 2022, using a Cessna aircraft with Leica RGBNIR camera. Flying altitude varied from 800m to 900m asl, which enabled production of orthomosaics at 10cm spatial resolution. Survey covered varying environments, illumination conditions, livestock sizes and grazing behaviour, which are typical challenges for livestock detection (Fig. 1).

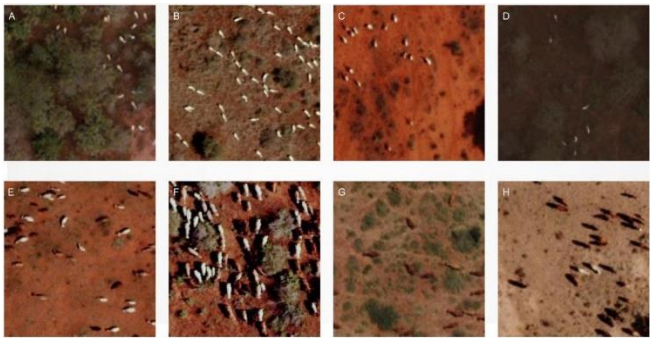

Fig. 1. Examples of livestock with diverse background and other challenges of livestock detection: a) dense shrubland, b) grassland, c) bare land, d) poorly illuminated image, e) sparsely distributed livestock, f) livestock in herds, g) brown fur livestock and h) presence of shadows.

2.3. YOLOv5

YOLOv5 is the fastest one stage object detector. The model detects and classifies an object in a single pipeline operation rather than generating regions of interest as two-stage detectors. Currently, YOLOv5 is the most stable and popular YOLO version. It is composed of five architectures; nano (n), small (s), medium (m), large (l), and extra-large (x), which differ based on the depth of architecture. YOLOv5 has three sections: the backbone made up of Darknet-53, which is responsible for feature extraction, the neck, which consists of CSPNET for fine tuning, and the head layer, which predicts by generating bounding boxes, assigns confidence scores around the detected object and classifies the target. [18]

2.4. Livestock annotation and training datasets

We split the image mosaics into 4km^2 tiles, which were visually scanned for the presence of livestock, before being cropped into small patches of 200 x 200 pixels. Use of small patches minimizes detection errors and improves the computation efficiency. Furthermore, we screened the small patches to retain only the ones with livestock present, resulting in 377 patches. These patches were manually annotated using MakeSense online platform (https://www.makesense.ai/) to draw bounding boxes around the target and to assign the livestock labels to a single class 'cattle'. The patches and annotations were split in the ratio of 70:15:15 into training, validation and testing sets, respectively.

In addition to original imagery, we used 300 background patches (non-cattle) and an augmented dataset for comparison (Table 1). The augmentation strategies included rotation, flip, random crop, and adjustments of saturation, brightness, and contrast. Augmentation was aimed at increasing the number of training data and make the model more robust. Four different combinations of datasets were trained on YOLOv5 architectures (Table 1).

Table 1. Four experiments to assess YOLOv5 architecture for aerial imagery.

	Experiments	#Images	#Annotations
Exp1	Original data	327	2440
Exp2	Original + background dataset	627	2440
Exp3	Original + augmented dataset	3146	26917
Exp4	Original + augmented + background dataset	3446	26917

To the test the ability of the models trained with aerial imagery to generalize on coarser resolution VHR satellite imagery, we generated a synthetic test dataset by downsampling the original test aerial imagery dataset to 30cm resolution by cubic convolution resampling technique. This spatial resolution was selected as it resembles the spatial resolution of the highest VHR imagery currently on the market, Pleiades NEO (www.intelligence-airbusds.com).

2.5. Model training and validation

We trained the YOLOv5 architectures at 300 epochs with the initial weights obtained from MS COCO dataset [13]. We fine-tuned the hyperparameters using generic algorithm applied to the original dataset (Exp1 in Table 1). The best fit parameters were found after 255 generation of 10 epochs and batch size of 16. The models were implemented on NVIDIA Volta V100 GPU. We used Stochastic Gradient Descent

doi:10.2760/46796

(SGD) for model optimization for all the models, with the fine-tuned learning rate 0.01636, and applied early stopping for regularization against overfitting. The models were evaluated based on Precision (P), Recall (R), Average Precision (AP@0.5) and mean average precision (mAP, 0.5:0.95). mAP is the mean of average precisions between the intersection over union (IoU) of thresholds between 0.5 to 0.95 at an interval of 0.05. This is the standard evaluation metric when using COCO dataset for pretraining the weights.

3. RESULTS

The best mAP (50.2%) was produced by the inclusion of background and augmented data (Exp4), and use of YOLOv5x architecture (Table 2). For all the models, the accuracy was the lowest when using only the original data (Exp1). The use of augmentation strategies (Exp3) improved mAP substantially in each case (average increase in mAP was 10.8%) while the inclusion of background data (Exp 2) had only minor effect (average increase of 1.0%). Furthermore, the inclusion of background data in addition to the augmented data (Exp4) had negative impact for some of the architectures, as the mAP for YOLOv5m declined by 1% in Exp4 compared to Exp3.

Table 2. Comparison of YOLOv5 architectures mAP (0.5:0.95) on test dataset.

Model	Exp1 (%)	Exp2 (%)	Exp3 (%)	Exp4 (%)
YOLOv5s	40.5	41.0	50.5	50.7
YOLOv5m	40.3	42.1	51.6	50.6
YOLOv5l	39.6	41.7	51.7	51.5
YOLOv5x	41.4	42.2	51.2	**51.8**

In estimating population counts, YOLOv5x in Exp4 had the least error in total count (-5.6%), detecting 454 cattle compared to the 481 manually annotated labels (Table 3). Inclusion of augmentation strategies (Exp3 and Exp4) reduced the underestimation by 10.4% and 10.5%, respectively. The decline in count error when including augmented data were considerably smaller for YOLOv5s in comparison to other architectures.

Table 3. Error in population count estimates detected against the manual labels.

Model	Exp1 (%)	Exp2(%)	Exp3(%)	Exp4(%)
YOLOv5s	-21.2	-21.2	-11.6	-16.8
YOLOv5m	-16.6	-21.6	-8.9	-7.3
YOLOv5l	-22.7	-9.8	-8.7	-7.7
YOLOv5x	-18.7	-21.0	-8.3	**-5.6**
Average	-19.8	-18.4	-9.4	-9.3

The large error in counts in Exp1 and Exp2 were due to undetected cattle when there is crowding of herds (Fig. 2). Several cattle were not detected in Exp1 and Exp2 (Fig 2a and Fig 2b), while minimal or zero misdetection were observed in Exp3 and Exp4 (Fig. 2c and Fig. 2d).

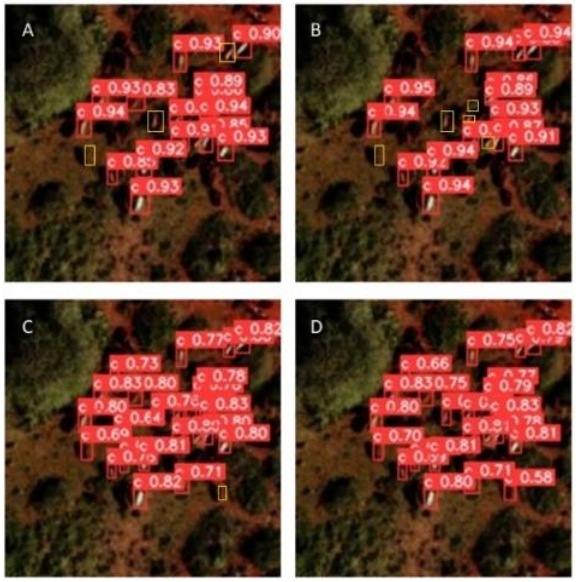

Fig. 2. Detected cattle and their confidence probabilities in red boxes while undetected cattle shown in yellow boxes in a) original data, b) inclusion of background data, c) inclusion of augmented dataset and d) inclusion of both augmented and background dataset.

Finally, YOLOv5x with the best mAP trained in Exp4 was tested on simulated VHR satellite imagery. The mAP and the estimated counts decreased by 5.5% and 23.3%, respectively (Table 4).

Table 4. Evaluation metrics of the simulated satellite images datasets.

#Images	#Labels	Precision (%)	Recall (%)	mAP (%)	Counts (%)
57	441	96.3	70.3	46.3	-28.9

4. DISCUSSION

The results show that low altitude aerial imagery and YOLOv5 one stage object detector can be used for training a model to detect medium to small size livestock in African rangelands with variable and heterogenous background. The best model had acceptable accuracy in terms of total population count although underestimation was observed in case of denser cattle herds. Furthermore, impact of augmentation strategies collaborates finding in prior studies [14], [15], on improving the detection accuracy by making the models more robust to detect cattle in different backgrounds. While inclusion of background dataset was aimed at reducing false positives, it had relatively small positive impact on the accuracy and estimated counts.

In addition, we showed that it is possible to upscale a YOLOv5x to detect medium to small size livestock on simulated VHR satellite imagery. The high depth of parameters and networks of YOLOv5x [16], coupled with the diverse and sufficient training sample enables the model to

doi:10.2760/46796

detect small objects. However, lower spatial resolution increases image blurriness, which reduces detection to distinguish brown and dark coloured cattle breeds from the background. This led to considerable underestimation of cattle population number compared to higher resolution aerial imagery.

The results of this study are based on simulated satellite images, which enabled comparison of counts based on higher resolution aerial imagery and VHR satellite image. In future, testing the model on real VHR satellite images with varying from various sources and at spatial resolution <0.5m will be essential. Furthermore, comparison of the YOLOv5 model with other point-based detection and segmentation architectures in different landscapes are needed.

5. ACKNOWLEDGEMENT

This work is funded by European Union DG International Partnerships under DeSIRA program (FOOD/2020/418-132) through the ESSA project. The contents of this document are the sole responsibility of the authors and can under no circumstance be regarded as reflecting the position of European Union.

REFERENCES

[1] R. Kariuki, S. Willcock, and R. Marchant, "Rangeland livelihood strategies under varying climate regimes: Model insights from southern Kenya," *Land (Basel)*, vol. 7, no. 2, 2018, doi: 10.3390/land7020047.

[2] J. C. van Gemert, C. R. Verschoor, P. Mettes, K. Epema, L. P. Koh, and S. Wich, "Nature conservation Drones for automatic localization and counting of animals. In : Agapito L., Bronstein M., Rother C. (eds) Computer Vision," in *ECCV 2014 Workshops*, Springer Verlag, 2015, p. VI. doi: 10.1007/978-3-319-16178-5.

[3] A. Delplanque, S. Foucher, P. Lejeune, J. Linchant, and J. Théau, "Multispecies detection and identification of African mammals in aerial imagery using convolutional neural networks," *Remote Sens Ecol Conserv*, vol. 8, no. 2, pp. 166–179, Apr. 2022, doi: 10.1002/rse2.234.

[4] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, IEEE Computer Society, Dec. 2016, pp. 779–788. doi: 10.1109/CVPR.2016.91.

[5] W. Liu *et al.*, "SSD: Single Shot MultiBox Detector," in *Eccv*, B. Leibe, Ed., Springer International Publishing, Dec. 2016, pp. 21–37. doi: 10.1007/978-3-319-46448-0_2.

[6] R. Girshick, "Fast R-CNN," *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, doi: 10.1109/ICCV.2015.169.

[7] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2014, pp. 580–587. doi: 10.1109/CVPR.2014.81.

[8] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," *IEEE Trans Pattern Anal Mach Intell*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017, doi: 10.1109/TPAMI.2016.2577031.

[9] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask R-CNN," *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2017-Octob, pp. 2980–2988, 2017, doi: 10.1109/ICCV.2017.322.

[10] F. Yang, N. Zhu, S. Pei, and I. Cheng, "Real-time open field cattle monitoring by drone: A 3D visualization approach," in *International Conference on Computer Graphics, Visualization, Computer Vision and Image Processing*, 2021, pp. 124–128.

[11] J. G. A. Barbedo, L. V. Koenigkan, P. M. Santos, and A. R. B. Ribeiro, "Counting cattle in UAV images-dealing with clustered animals and animal/background contrast changes," *Sensors (Switzerland)*, vol. 20, no. 7, Apr. 2020, doi: 10.3390/s20072126.

[12] F. Sarwar, A. Griffin, S. U. Rehman, and T. Pasang, "Detecting sheep in UAV images," *Comput Electron Agric*, vol. 187, Aug. 2021, doi: 10.1016/j.compag.2021.106219.

[13] T. Y. Lin *et al.*, "Microsoft COCO: Common Objects in Context," in *European Conference on Computer Vision-ECCV 2014 Part of the Lecture Notes in Computer Science*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds., Cham: Springer International Publishing, 2014, pp. 740–755. doi: 10.1007/978-3-319-10602-1_48.

[14] J. Brown, Y. Qiao, C. Clark, S. Lomax, K. Rafique, and S. Sukkarieh, "Automated aerial animal detection when spatial resolution conditions are varied," *Comput Electron Agric*, vol. 193, no. 106689, 2022, doi: 10.1016/j.compag.2022.106689.

[15] D. Wan, R. Lu, S. Wang, S. Shen, T. Xu, and X. Lang, "YOLO-HR: Improved YOLOv5 for Object Detection in High-Resolution Optical Remote Sensing Images," *Remote Sens (Basel)*, vol. 15, no. 3, p. 614, 2023, doi: 10.3390/rs15030614.

[16] R. Baidya and H. Jeong, "YOLOv5 with ConvMixer Prediction Heads for Precise Object Detection in Drone Imagery," *Sensors*, vol. 22, no. 21, pp. 1–17, 2022, doi: 10.3390/s22218424.