Computers and Electronics in Agriculture 171 (2020) 105300

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

journal homepage: www.elsevier.com/locate/compag

Automated cattle counting using Mask R-CNN in quadcopter vision system

Beibei Xu[a], Wensheng Wang[a,*], Greg Falzon[b,c], Paul Kwan[d], Leifeng Guo[a], Guipeng Chen[e], Amy Tait[f], Derek Schneider[b,c]

[a] Agricultural Information Institute, Chinese Academy of Agriculture Sciences, Beijing 100086, China
[b] School of Science & Technology, University of New England, Armidale, NSW 2351, Australia
[c] Precision Agriculture Research Group, University of New England, Armidale NSW 2351, Australia
[d] School of Information Technology & Engineering, Melbourne Institute of Technology, Melbourne, VIC 3000, Australia
[e] Agricultural Economics and Information Institute, Jiangxi Academy of Agriculture Sciences, Nanchang 330200, China
[f] School of Environmental and Rural Science, University of New England, Armidale, NSW 2351, Australia

ARTICLE INFO

Keywords:
Object detection
Deep learning
Remote monitoring
Livestock management
Quadcopter vision system

ABSTRACT

The accurate and reliable counting of animals in quadcopter acquired imagery is one of the most promising but challenging tasks in intelligent livestock management in the future. In this paper we demonstrate the application of the cutting-edge instance segmentation framework, Mask R-CNN, in the context of cattle counting in different situations such as extensive production pastures and also in intensive housing such as feedlots. The optimal IoU threshold (0.5) and the full-appearance detection for the algorithm in this study are verified through performance evaluation. Experimental results in this research show the framework's potential to perform reliably in offline quadcopter vision systems with an accuracy of 94% in counting cattle on pastures and 92% in feedlots. Compared with the existing typical competing algorithms, Mask R-CNN outperforms both in the counting accuracy and average precision especially on the datasets with occlusion and overlapping. Our research shows promising steps towards the incorporation of artificial intelligence using quadcopters for enhanced management of animals.

1. Introduction

Animal husbandry accounts for a large proportion of agriculture in many agricultural developed countries. In order to meet the increasing population's demand for meat and to respond to changes in people's dietary habits, there is a definite need for improving livestock production and welfare (Liaghat and Balasundram, 2010). The management for livestock is developing from small-scale and subsistence farming towards intensive and specialized grazing. Complicating factors such as lack of labour, difficulties in real-time monitoring and high costs in management have presented serious challenges to the large-scale and intensive pasture-based production systems. This requires precise and cost-effective technology methods to address these challenges in animal agricultural systems.

1.1. Animal remote monitoring

Recent advancements in information technology for remote monitoring have enabled farmers to obtain more accurate

and valuable information about an animals' behaviour and the environment in which they live to improve meat quality, maximize production and promote animal health and welfare (Ruiz-Garcia et al., 2009). Wireless Sensor Technologies which can assist in providing continuous and remote monitoring information in real time, have brought great changes to information perception, making remote monitoring and management possible (Ruiz-Garcia et al., 2009). Wearable technologies such as RFID, Accelerometer Sensors, GPS Collars and Smart Ear Tags are already available for farmers to monitor behaviour and movement, body temperature, heart rate and other physiological factors to avoid morbidity and mortality of animals, and thereby reduce production losses (Frost et al., 1997; Handcock et al., 2009; Marsh et al., 2008; Neethirajan, 2017; Neethirajan et al., 2017; Ruiz-Garcia et al., 2009; Sellier et al., 2014; Van Nuffel et al., 2015). Motion-activated cameras (camera traps) have also been used as a cost-effective approach to recording an animals' presence, location and activity (Yu et al., 2013). Animal species can also be identified and counted automatically based on camera trapping imagery (Norouzzadeh et al., 2018). In the diagnosis of animal diseases and detection of habitats, Thermal Infrared Imaging has played an increasingly indispensable part in detecting early inflammations of limbs

* Corresponding author.
E-mail address: wangwensheng@caas.cn (W. Wang).

https://doi.org/10.1016/j.compag.2020.105300
Received 14 October 2019; Received in revised form 15 February 2020; Accepted 18 February 2020

and follow up pregnancy in animals (Gonzalez et al., 2016; Handcock et al., 2009).

In sparsely populated countries with existing advanced animal husbandry techniques such as Australia and New Zealand, due to vast areas and distances in the rangelands, it remains a critical but challenging task to detect animals and obtain accurate information on an individual basis without delay especially in extreme weather conditions. Real-time detection can avoid loss or theft of animals as well as prevent animal incursions from other farms resulting in overgrazing of pastures. A significant limitation of existing ground-based monitoring techniques such as smart ear tags, camera traps and Infrared Thermal Imaging is a result of the relatively large geographic scope and complex terrain where animals could possibly be identified and tracked (Norouzzadeh et al., 2018). Remote sensed imagery could be utilised as a potential alternative to ground-based animal surveys. The use of Quadcopters as an emerging but promising approach will combine with machine learning algorithms to revolutionize livestock management. Compared with other technologies, quadcopters can: (1) complete low-altitude and ultra-low-altitude flight paths; (2) obtain high-resolution images at any time over a wide range of weather conditions; (3) acquire images quickly in small areas and over inaccessible rugged terrain. Even so, the accurate and reliable counting of animals in quadcopter acquired imagery is one of the most important but challenging tasks in intelligent livestock management (Barbedo and Koenigkan, 2018). The challenge for quadcopters to become a truly efficient livestock monitoring tool are image processing algorithms matching the corresponding functions (Gonzalez et al., 2016). Therefore, our work focuses on algorithms for animal detection and counting to automatically conduct population censuses of cattle in the images captured by the quadcopter and to further demonstrate the potential for quadcopter machine vision system learning applied in livestock management (Chamoso et al., 2014; Handcock et al., 2009; Norouzzadeh et al., 2018).

1.2. Animal detection and counting using a quadcopter

Thanks to recent developments in machine vision, the application of quadcopters in animal detection and counting has increased (Chamoso et al., 2014). Examples of applying UAV for population estimation include birds (Abd-Elrahman et al., 2005; Chabot and Francis, 2016; Descamps et al., 2011; Grenzdörffer, 2013), mammals, (Hodgson et al., 2013; Koski et al., 2009; Rey et al., 2017; Vermeulen et al., 2013), wildlife (Chabot, 2009; Chabot and Bird, 2015; Chrétien et al., 2015; Gonzalez et al., 2016; Lhoest et al., 2015) and livestock (Chamoso et al., 2014; Hollings et al., 2018; Kellenberger et al., 2018; Longmore et al., 2017; Van Nuffel et al., 2015). The most straightforward method for automated techniques in detecting and counting animals is image segmentation which analyses the individual pixels in the UAV images with a specified spectral threshold (Chabot and Francis, 2016; Mejias et al., 2013). This works best in the situation where animals contrast sharply with their background. As for more complex scenes which need to consider texture, colour, spatial context and so on, Abd-Elrahman et al (2005) improved a template matching approach in conjunction with spectral characteristics to carry out multi-stage pattern recognition and counting in the UAV equipped with video imaging sensors (Abd-Elrahman et al., 2005). Gonzalez et al (2016) integrated thermal image capabilities into a UAV and proposed to detect, classify and track wildlife to obtain promising estimate within the area surveyed. They used pixel intensity threshold and template matching binary mask respectively in different cases (Gonzalez et al., 2016).

Nonetheless, the resulting estimate seems to be unsatisfactory if animals overlap seriously in the images. Such a situation is likely to occur in the livestock monitoring context where imagery is acquired of herds of animals. The use of machine vision to detect livestock from UAV imagery has been demonstrated as successful and promising for further research (Sadgrove et al., 2018). Indeed, a wide range of machine learning approaches have been explored for the task of animal detection and counting from remotely sensed imagery. Descamps et al. (2011) employed the method of unsupervised learning to achieve clustering and automatic counting among birds according to shape and spectral discreteness. Traditional supervised multispectral image classification (Chabot and Francis, 2016) such as Maximum Likelihood Classification based on pixels with ArcGIS's Spatial Analyst extension was also introduced to identify and count animals. However, species classes in some images needed to be labelled manually prior to the training examples which required user's knowledge to accurately label animals (Grenzdörffer, 2013) and the quality of training data largely determined the effect of supervised method. Unsupervised classification has also been demonstrated overestimating populations (Hollings et al., 2018). Chrétien et al (2015) combined object-based image analysis with spectral and spatial information for white-tailed deer detection in multispectral imagery, and also for multispecies detection and counting in a controlled environment. Compared with supervised and unsupervised pixel- based image classification approaches, they performed more accurately but slow in processing and also required specialist knowledge (Chrétien et al., 2015).

The Convolutional Neural Network (CNN) has also been considered as a practical detection and counting technique with regard to variable inputs, processing speed and accuracy for image recognition (Barbedo and Koenigkan, 2018; Chamoso et al., 2014). Chamoso et al. (2014) combined CNN with UAV system to keep track of counting animals detected in video recordings taken from UAV. Maire et al (2014) applied CNN in aerial imagery and proposed an approach for training negative example-selection, which proved very effective for automatically annotating and detecting dugongs. Whereas, the excellent performance of convolutional neural networks used in animal detection heavily depend on the large datasets including positive and negative examples. Kellenberger et al (2018) presented a solution to sparse data in the aerial images including class weights application to reduce the impact of complex background using curriculum learning (Bengio et al., 2009) to strengthen the feature training of animals and backgrounds. In recent years, deep convolutional neural networks such as R-CNN (Girshick et al., 2014), Fast R-CNN(Girshick, 2015), Faster R-CNN (Ren et al., 2015) and Mask R-CNN (He et al., 2017) have shown their great potential in object detection and classification of thousands of global images due to higher accuracy, precision and a quicker processing speed. They have been used to detect and count fruit (Sa et al., 2016; Stein et al., 2016) as well as assess the quality of fruit automatically (Jail et al., 2018) which achieved good results.

For animal detection, as with fruit detection, consideration should be given to changes in illumination, overlapping, and small differences between background and target objects. Faster R-CNN has been applied to the task of detection and individually identifying 30 Holstein-Friesian cattle from DJI MkI quadcopter videos (Andrew, 2017a). The Faster R-CNN algorithm proposed produced an excellent detection and localisation performance of 99.3% in a relevant dairy production setting (Andrew, 2017a). Although the results were impressive, the scenario examined was very specific and arguably, which was one of less challenging computer vision scenarios for object detection and localisation consisting of cattle with distinctive black and white coat patterns contrast with lush green pasture. In contrast, there are many more diverse and challenging livestock monitoring scenarios including visual clutter (vegetation and other natural elements), strong lighting contrast and shadows (from farm infrastructure) and high density (tightly packed herds or stock constrained in feedlots). There is a need to perform wider assessment of cattle counting algorithm performance across a range of livestock production settings.

Of all of the approaches proposed in the literature, the Mask R-CNN appears to the most promising for the monitoring of cattle. The Mask R-CNN approach allows both counting the number of stock in the image and also identification/extraction of the pixels associated with each individual animal. Such extraction leads to further applications in the

machine vision pipeline, e.g. biometrics and welfare monitoring. The focus of this paper is on counting but these other tasks are also noted and a motivation for the use of Mask R-CNN. Mask R-CNN has also been demonstrated to be robust to illumination, deals adequately with large numbers of over-lapping and close proximity objects. Mask R-CNN can also detect objects with similar texture or colours to background objects (He et al., 2017). Despite the general appeal of Mask R-CNN (Qiao et al., 2019), it has not been evaluated in great detail for precision livestock monitoring applications using quadcopters. Given the urgent need to develop technologies which can assist with livestock production and welfare management, it is timely to assess the application of a state-of-the-art machine learning algorithm for precision livestock monitoring. In this work, we explore the application of machine learning in cattle detection and counting using the cutting-edge instance segmentation framework, Mask R-CNN, a stronger robust method, aiming to build a quadcopter machine vision system for monitoring livestock in a precise and effective way.

2. Related work

Object detection is a fundamental task in the field of computer vision, which is aimed to accurately find the target objects and their location in the images. In the case of multiple objects, the processing pipeline can be further extended to classify the images into known classes (Radovic et al., 2017). Cattle, due to their major importance to the livestock industry, were selected as the case-study to explore the performance of Mask R-CNN based object detection within quadcopter imagery. In traditional object detection, a sliding-window frame is generally adopted with three stages. First, sliding windows of different sizes are used to generate the region proposals, followed by extracting visual features through models such as Haar-like feature and Histogram of Oriented Gradient feature (HOG) (Dalal and Triggs, 2005). The final stage is to put the features selected into the classifier for identification such as Support Vector Machine model (SVM). For instance, Viola-Jones algorithm based on Haar-features and Adaboost boosting algorithm was implemented to detect black-backed jackal faces in images (Pathare, 2015), and the Local Binary Pattern (LBP) adopting AdaBoost algorithm was used to detect dangerous animals including moose, elk and cow (Zhou, 2014). However, traditional object detection is often not on target while selecting the region proposals, rendering the time complexity high and many of the windows redundant. In addition, manually-designed features in the traditional object detection are not robust enough to deal with wide diversity image changes encountered in practice.

In contrast, more recent objection approaches combine artificial neural network and deep learning technology via the convolutional neural network, which has combined local region perception, feature extraction along with a classification process to train the network globally. The weight-sharing network structure is invariant to the translation, tilt, zoom, or other forms of deformation of images, so that images can be directly used as the input of the network, avoiding the extraction of complex features and their reconstruction process in the traditional object detection (Zhang, 1988; Zhang et al., 1990). Among the state-of-the-art object detection algorithms in the field of deep learning, the algorithm based on region proposals such as R-CNN (Girshick et al., 2014), Fast R-CNN (Girshick, 2015), Faster R-CNN (Ren et al., 2015) and Mask R-CNN (He et al., 2017), and the algorithm based on regression such as YOLO (Redmon et al., 2016) and SSD (Liu et al., 2016) has achieved the best performance in terms of mean average precision (mAP) and frame per second (FPS). Importantly, these deep learning-based object detectors have been demonstrated to significantly out-perform traditional based object detection algorithms (Lee, 2015).

The regression-based algorithms such as YOLO and SSD requires the generation of some regions of interest according to feature extraction firstly, then classification of every region and finally produce a bounding-box regression. This provides an end-to-end process for the regression-based detection with direct prediction of target classification and a bounding-box using a single neural network. Despite fast speed and applicability of real-time detection, regression-based detectors (such as SSD and YOLO) have been demonstrated to achieve lower mean average precision (mAP) especially for low image resolution as compared to other detectors such as Faster R-CNN (Huang et al., 2017; Redmon et al., 2016; Redmon and Farhadi, 2018). In addition, YOLO also has difficulty detecting small objects as well as overlapping objects, and is hard to distinguish objects with a similar color to the background (Burić et al., 2018; Sommer et al., 2018), which are very common in the outdoor type of farmlands. The detection of objects in imagery with low resolution is very relevant for the monitoring of livestock using quadcopters, as monitoring stock from a distance is more efficient and less intrusive but produces smaller sized objects in the images. Because Faster R-CNN involves global average pooling in order to reduce the computation of first fully connected layer, this reduces the precision of spatial localization (Li et al., 2017). Instead, Mask R-CNN improves the RoI Pooling using RoIAlign to remove the harsh quantization of RoI Pooling, properly aligning the extracted features with the input to improve the accuracy of prediction (He et al., 2017; Li et al., 2017).

Therefore, Mask R-CNN (an extension of Faster R-CNN) which also allows for instance segmentation (associating specific image pixels to the detected object) is selected for further study. Instance segmentation allows not only the detection of each animal but also the delineation of its boundaries within the image thereby allowing further potential applications for livestock welfare monitoring. The benefits provided by instance segmentation allow for diverse future applications including estimation of animal pose and direction of travel. In this work however, we constrain interest to the object detector capabilities of Mask R-CNN. He et al. (2017) compared the detection performance of Mask R-CNN (ignoring the mask output) to Faster R-CNN and found slight increases in detection performance (+3.6 of average precision) (He et al., 2017). Based on the results reported by (He et al., 2017) and the potential future user afforded by instance segmentation, this paper aimed to examine the effectiveness of Mask R-CNN for the detection and counting of cattle in quadcopter imagery.

Common difficulties in livestock imagery such as diversities in cattle pose, heavy occlusions among a herd of cattle and repeat count for single cattle cropped into multiple images have brought great challenges to cattle detection and counting (Van Nuffel et al., 2015). Head detection, as typically applied in people detection (Gao et al., 2016; Van Nuffel et al., 2015) is introduced to recognize cattle in this paper and compare performance with full-appearance detection to decide which approach is best for detecting cattle in quadcopter imagery.

3. Materials and methods

3.1. Overview of our framework

The section describes the pipeline which is proposed for processing RGB images that are captured by a quadcopter to detect and count cattle using deep learning algorithm. The structure of cattle detection and counting in aerial images is illustrated in Fig. 1. The RGB image acquired by the drone is used to extract the feature from the full image using the convolution layers, and then the obtained feature map is sent to the Region Proposal Network (RPN) to generate Region of Interests (ROIs). The RoiAlign layer selects the feature corresponding to each ROI on the feature map according to the output of the RPN, and send them to the fully connected layer for classification prediction, mask prediction and bounding-box prediction. Ground truth was annotated manually for every cattle in the training sets and then network training was performed after labelling for parameters optimization, followed by cattle detection and counting in testing sets.

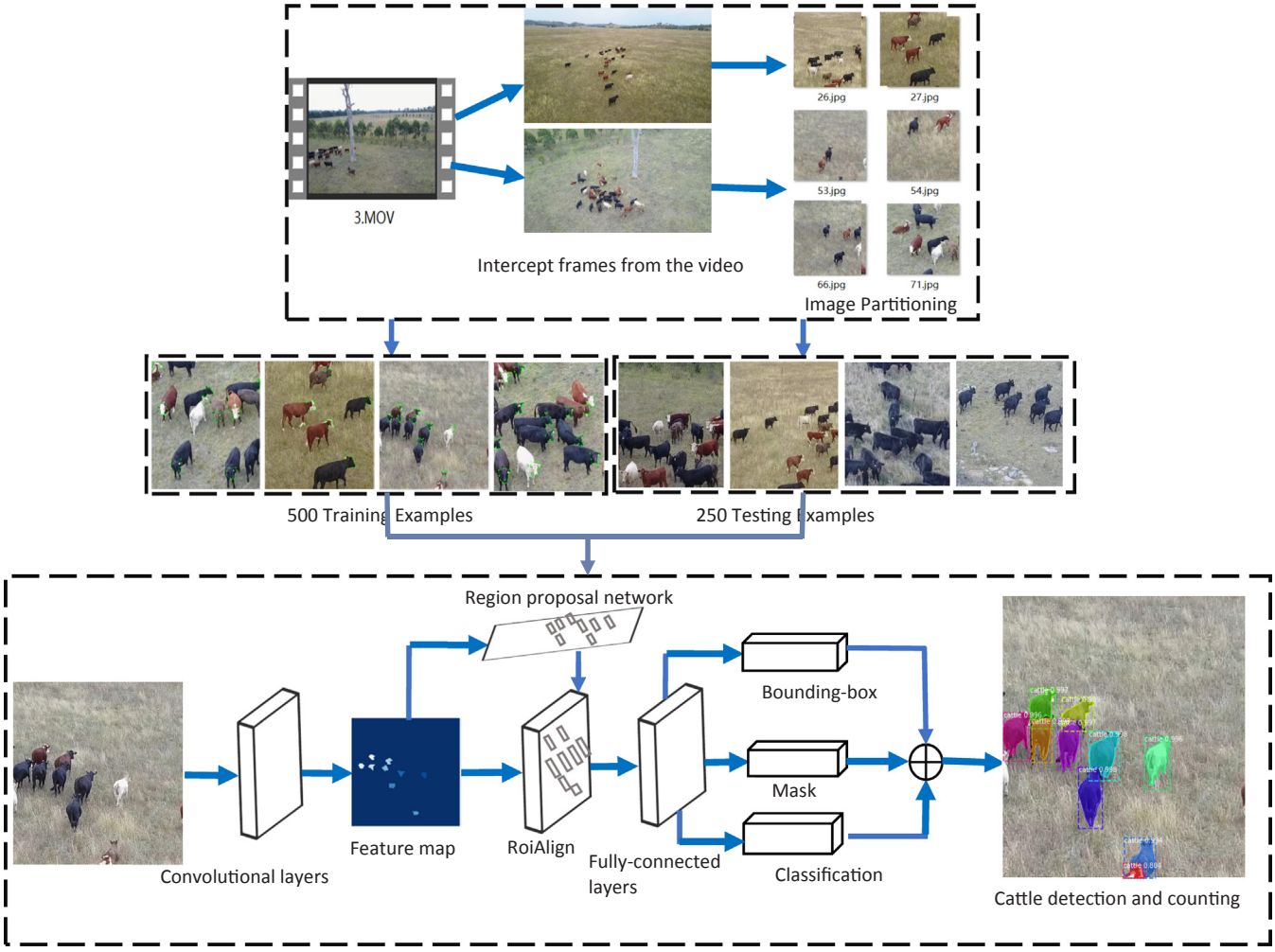

Fig. 1. The structure of the cattle detection and counting algorithm.

3.2. Datasets preparation and preprocessing

The limitation to effectively detect and count cattle using machine vision is the lack of suitable publicly available datasets because the published datasets like FriesianCattle dataset in (Andrew, 2017) are only one or two cows per image, little difference between images and other disadvantages discussed in Section 1.2. Therefore, we used a drone to collect representative image data sets both in extensive pasture and in feedlot environments separately. The datasets utilized in this research were collected from the Tullimba Research Feedlot (AEC18-038) owned by the University of New England, New South Wales, Australia and surrounding farmlands (AEC19-009) across seasons from Summer to Spring (February to October). Examples of cattle in different scenes are on display in Fig. 2. Three farmlands and one feedlot were chosen for the potential application of this technology in different cases and in different weather and backgrounds. Fifteen flight campaigns were conducted by the MAVIC PRO drone which is equipped with an integrated PTZ camera shown in Fig. 3. The camera has a 1/2.3-inch CMOS image sensor that can rotate flexibly both laterally and vertically. 4 K HD videos and 12-megapixel photos are captured by this stabilized camera. Considering the pixels and the delay of taking photos as well as the convenience of operation, videos were adopted to obtain the cattle datasets in different scenes at 30 frames per second in recordings and were saved in MOV format. The cattle datasets were captured at a height of 8–25 m with angles of inclination in the pastures and at a downward angle vertically in the feedlot to simulate the aerial detection (Sadgrove et al., 2017).

Fig. 2. Examples of cattle in the farmland (left) and in the feedlot (right).

The original images cropped from videos were in JPG formats at 4096 by 2160 pixels. In order to avoid overfitting when training the network and improve the processing speed, after extracting valuable data frames of every video in MATLAB, the selected images were clipped automatically using MATLAB to the size 512 * 512 pixels from the pastures and 1280 * 1280 pixels from the feedlot. The image size must be divisible by 2 at least 6 times to avoid fractions when downscaling and upscaling in the CNN algorithm. Each of the datasets both in the pastures and feedlot contain a total of 750 images consisting of 500 images for training and 250 images for testing. The dataset chosen for

Fig. 3. MAVIC PRO drone.

training and testing are from different key frames during cattle movement, and the ratio of training and testing under same environmental conditions is 2:1.

The specifications of dataset used in the experiments are displayed in the Table 1. Due to the limited space for each pen in the feedlot and the diverse body positions of the cattle, only data collection for counting, was obtained vertically using the quadcopter. But the head of cattle is very difficult to distinguish from other parts of the body even with our eyes because of the small size of head compared to the body and similar colour between the background of the image and the cattle. So, only the full-appearance is performed in the feedlot. In addition, the angles and heights of data collection in the feedlot and the pasture are different, so the size of cattle in the original images differs greatly. In order to reduce the deviation caused by varying visual scales of cattle in the images and to be within computational constraints, image dimensions of 1280 square pixels and 512 square pixels were selected for the feedlot and pasture environments respectively.

The publicly available image annotation tool known as LabelMe (Russell et al., 2008) is used to label the ground truth for head and full-appearance respectively using polygon for training datasets (see Fig. 4). For labelling, points are clicked along the outside edge of every cattle in the images until connected to the starting point. Then the class label named cattle needs to be marked on the bubble pop up on the screen. The ground truth data was stored in a table format aligned with that required by the Mask R-CNN framework for data annotation.

3.3. The detection and counting algorithm

Extended from Faster R-CNN, Mask R-CNN additionally provides a mask prediction branch composed of a small Fully Convolutional Network for segmenting each Region of Interest (ROI) with simultaneous classification prediction and bounding-box prediction. Same with Faster R-CNN, the object detection of Mask R- CNN is also divided into two stages: (i) Region Proposal Network (RPN) and (ii) Classification based on binary mask. RPN is a newly high-sufficient proposal generation network in the Faster R-CNN which replaces the selective search method in the previous RCNN and Fast R-CNN.

However, Mask R-CNN additionally produces a binary mask besides the class label and bounding-box for each ROI (He et al., 2017).

Table 1
Dataset specifications from pasture and feedlot environments.

Case	Description	Training set	Test set	Image pixels
Pasture	Full-appearance detection	500	250	512 * 512
Pasture	Head detection	500	250	512 * 512
Feedlot	Full-appearance detection	500	250	1280 * 1280

Classification prediction in Mask R-CNN is closely related with mask branch, so the mask is also exploited to get the spatial structure of an object through the pixel-to-pixel alignment in the convolutional layers being encoded. Considering the potential regional misalignment between the input and extracted feature map with no impact on ROI Pooling, RoiAlign is adopted in the Mask R-CNN using the bilinear interpolation to improve the precision of the model. The details for Mask R-CNN evolved from CNN and other region-based approaches are expounded in the original papers (Girshick, 2015; Girshick et al., 2014; He et al., 2017; Ren et al., 2015) and this paper mainly describes the key procedures in the application of the algorithm.

- Region Proposal Network

On the feature maps from the convolutional layers, the network performs convolution operation on a 3 * 3 pixel sliding window. For each centre point in the feature map, k anchors with different scales and aspect ratios are selected and then are mapped on the original feature maps according to the scales and aspect ratios, producing thousands of region proposals. Each point in the feature maps generates feature codes for the corresponding window regions which is corresponded to the low-dimensional feature codes of 512 dimensions in Mask R-CNN. Then, the low-dimensional feature codes are performed by a $1 \times 11 \times 1$ convolution operation, which outputs $2 * k$ classification features and $4 * k$ regression features, respectively corresponding to the confidence scores and the relative coordinates of anchors. Based on the ranking of classification scores, the first 2000 regression feature boxes are selected and the values of relative coordinates are decoded into the absolute coordinates via the following formulas (1) and (2):

$$t_x = (x - x_a)/w_a,\ t_y = (y - y_a)/h_a \tag{1}$$

$$t_w = \log(w/w_a),\ t_h = (h/h_a) \tag{2}$$

Here (x_a, y_a) is the coordinates of the centre of the anchor and (w_a, h_a) is the height and width of the anchor. (x, y) is the coordinate of the centre of the predicted ROI in the original image and (w, h) is the height and width of the ROI predicted in the original image. (t_x, t_y) is the regression value of the coordinates of the centre on the feature map and (t_w, t_h) is the regression value of the height and width on the feature map. A certain number of Region of Interests (200 in Mask R-CNN) are then selected to be trained through Non-maximum suppression which needs to compare the ROI with ground truth. Specifically, if the value of intersection-over-union (IoU) between the predicted bounding boxes in the ROI with ground truths is larger than a set threshold, there must be targets in this ROI which will be regarded as foreground and background otherwise.

- Loss Function

Multi-task loss function is used in the training for Mask R-CNN which consists of three parts: the classification loss of the bounding box, the position regression loss of bounding box and the loss of the mask following formulas.

$$\mathrm{L} = L_{cls} + L_{box} + L_{mask} \tag{3}$$

$$L_{cls} = -\log[p_i^* p_i + (1 - p_i^*)(1 - p_i)] \tag{4}$$

$$L_{box} = r(t_i - t_i^*) \tag{5}$$

$$L_{mask} = \mathrm{Sigmoid}(\mathrm{Cls}_k) \tag{6}$$

Here, p_i is the predicted probability for ROI in the classification lossL_{cls} andp_i^* for ground truth is 1 if the ROI is regarded as foreground or 0 otherwise.$t_i$is the vector of absolute coordinates for predicted bounding box (see formula 5) and t_i^* is for ground truth in the position regression loss of bounding box where r is the robust loss function to calculate regression error referring to (Girshick et al., 2014). Each ROI predicts

Fig. 4. Examples of annotations (green) for head detection in the first line and for full-appearance detection in the second line. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

an output of K*mˆ2 dimensions through the mask branch, and it encodes K binary masks with a resolution of m*m, corresponding to K classes. The loss of the maskL_{mask} is defined as the Average Binary Cross-entropy Loss which performs sigmoid function on each pixel in the ROI. For class k (Cls_k), the mask loss is shown in formula 6.

3.4. The implementation details and evaluation protocol

In view of relatively simple workflow and compatibility with other APIs, TensorFlow does not require any compiling time, which allows for faster iteration of models. Furthermore, the distributed architecture does not increase the time required for model training of large datasets. Multiple pre-trained machine learning models in TensorFlow are also available for use. Keras, an advanced neural network framework for Python which is added to TensorFlow to provide more modern APIs, allows import of the Resnet101 model and utilises data flow graphs to perform calculation. Therefore, the Mask R-CNN algorithm utilized in this paper is implemented on the TensorFlow framework. Python is selected as the programming language due to its code efficiency and comprehensive support for deep learning algorithms. Based on the Python 3.6 environment, the required toolkits such as 'numpy' and 'skimage' are obtained via anaconda3. The TensorFlow1.3 and Keras 2.15 that are compatible with the Python version were also installed.

Transfer learning can be defined as the tuning of an existing convolutional network to perform new tasks. It has become an essential part of machine learning as it provides a way to train a network when limited annotated data exists for the intended task. The network in this paper was initialized by a Resnet101-pre-trained model using COCO datasets (Ren et al., 2015) and the network head was used for multi-tasks of classification, bounding-box and mask prediction (He et al., 2016). To avoid destroying the extraction ability of convolutional layers, all the backbone layers were frozen and only the network head was trained independently using the training dataset of 500 aerial images in each case. The global layers were then fine-tuned to optimize the key parameters to achieve higher accuracy and faster processing speed. During network training, the loss of each ROI consisted of classification loss, bounding-box loss and mask loss, but the mask loss only exists in positive ROIs. So, the outcome was assigned positive if the Intersection-over-Union (IoU) between ROI and its ground-truth was at least a certain threshold otherwise it was classed as negative. Every image has some sampled ROIs with a 1:3 ratio of positive to negatives.

The Mask R-CNN implementation has been executed on a 64-bit version of Windows 10 laptop with Intel core i7-7560U CPU@2.4 GHz with 16 GB RAM. The Stochastic Gradient Descent (SGD) algorithm is adopted in network training with a weight decay of 0.001 and a momentum of 0.9 and an initial learning rate of 0.01. All the training experiments had a batch size of 50 images and the iterations of 3000. For the testing, all the testing results of each case were averaged 10 times from the dataset in the trained model with the same parameters. The number of proposals in the conv4_x was 1000 and the bounding-box prediction branch was performed on these proposals, then followed by non-maximum suppression to filter some overlapping proposals. The mask branch was run to the 200 detection boxes with highest scores so M masks can be predicted for per ROI but only the M-th mask is chosen (M is the predicted class by the classification predictor). The size of mask output was then adjusted to the size of ROI and binarized using a certain threshold ranging from 0 to 1.

In this paper, the precision, recall, F1 score and average precision (AP) are utilized as the evaluation metrics. The precision reflects the proportion of true predicted positive in all the predicted positive but the recall reflects the proportion of true predicted positive in all of the positives. For the precision-recall curve, the larger the area enclosed by the curve at different IoU threshold, the better the performance. F1 score is a statistical measure which is defined as the harmonic average between precision and recall, where it achieves the best performance at the value 1. IoU is the area of overlap between predicted and ground-truth bounding boxes divided by their area of union and represents the accuracy of the detection (formula 7).

$$\mathrm{IoU} = \frac{\text{detection result} \cap \text{ground truth}}{\text{detection result} \cup \text{ground truth}} \tag{7}$$

4. Experimental results

This section presents the performance evaluation of the proposed method to detect and count cattle on different experimental settings: (1) determine the optimal threshold of IoU; (2) compare the detection performance of cattle's head and cattle's full-appearance; (3) evaluate the effectiveness of proposed method by applying it in both pasture and feedlot situations; and (4) compare the proposed method with other

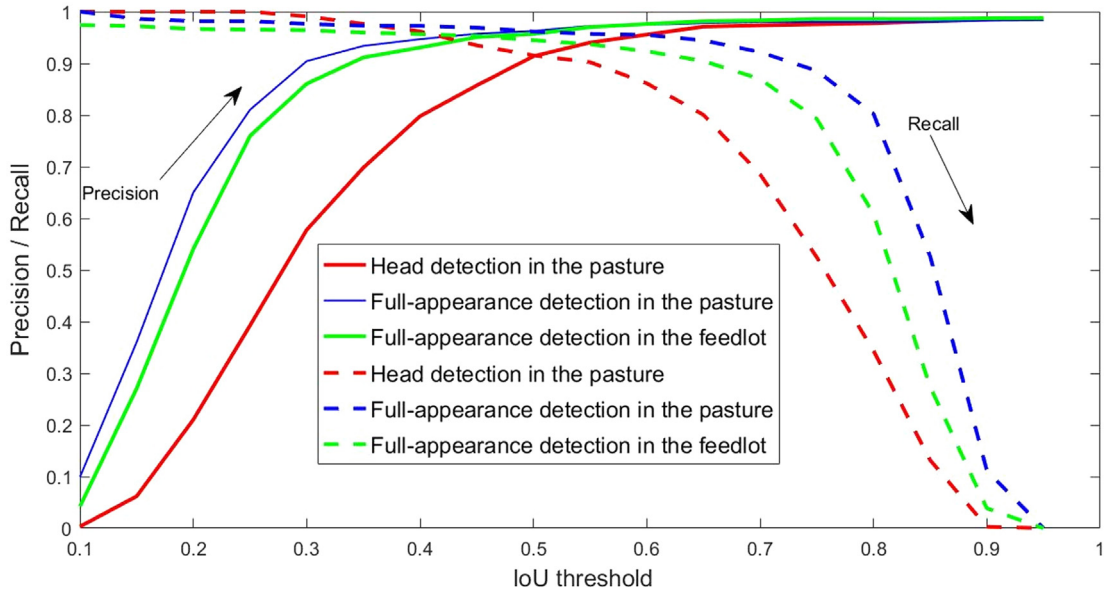

Fig. 5a. Comparisons of precision (solid lines) and recall (dotted lines) over different IoU thresholds for three detection cases.

state-of-the-art object detection algorithms.

4.1. Threshold selection of IoU

Intersection-over-Union (IoU) regarded as an important evaluation metric measures the overlap rate between the detection result predicted by the model and the ground truth. The threshold is a value between 0 and 1 and when it is reached, the dependent variable will change differently than before. The threshold value is critical to prediction performance of objection detection. If the threshold is set too large or too small, it will result in overlapping bounding-box predictions for the same cattle or missing cattle. In most cases, such as ImageNet challenge and PASCAL VOC challenge, the threshold from 0.4 to 0.7 is always chosen which is not necessarily the optimal threshold. This paper considers precision and F1 score as standard evaluation criterion in the single label detection (Fan and Lin, 2007) to evaluate variable thresholds under three different detection cases.

Fig. 5 shows that the precision, recall and F1 scores for three detection cases as a function of IoU thresholds. The solid lines in Fig. 5a represent the precision rates and the dotted lines for the recall rates. It can be observed that the precision increases but the recall decreases with the IoU threshold increasing. However, the precisions and the recalls in these cases have the same value while reaching a certain threshold (0.5) known as balance point which indicates that all the positive predictions are the true positives. Comparing the related F1 scores for three detection cases presented in Fig. 5b, resulted in the F1 scores being similar, where they get the best values at around that threshold. Obviously, at IoU = 0.5, the precision and the recall are high but not optimal but F1 score is maximised. F1 score is preferable as the metric for 'true positive detection' whilst precision is preferable for 'instance segmentation' (boundary extraction of each cow). Therefore, the performance of the threshold at 0.5 is significantly better than others. The best average precisions are achieved through the optimal threshold with 0.96, 0.92 and 0.94 for full-appearance detection, head detection in the pastures and full-appearance detection in the feedlot respectively.

4.2. Evaluation of detection and counting results

As previously mentioned in this section, the performance precision-recall curves whose IoU threshold ranges from 0.1 to 0.95. The results in Fig. 6 demonstrate the comparison between head detection and full-appearance detection in the pastures and performance for head detection in the feedlot. Due to the confined conditions and higher stock densities in the feedlot, it is very difficult to detect cattle's heads especially when they are in different body positions such as lying compared to standing. Consequently, the paper only performs on the full-appearance detection in the feedlot.

It can be seen from Fig. 6 there is an inverse relationship between the precision and recall which means the higher the precision the lower the recall. But we expect to detect all the target objects which means higher recall rates and also expected higher precision rates of the detected objects. At around recall = 0.91, 0.95, 0.96, the inflection point respectively appears in all three curves known as balance points where the precision and recall get best values, and then the precision drops sharply. Although three curves have some intersections, the precision of full-appearance detection in the pasture is higher than head detection and also than in the feedlot at balance points which is consistent with the conclusion in Section 4.1.

We compute the APs for bounding-box prediction masked as AP^{bb} and mask prediction as AP^{m} of three detection cases over a variety of IoU thresholds and F1 scores at the balance points shown in Table 2. These points imply that the predicted positives are all true positives. As observed, cattle counting based on full-appearance detection, yields an AP of 95% for bounding-box prediction, 94% for mask prediction and a F1 score of 0.96, which are both higher than head detection. In addition, the results concerning counting errors for all the test images in Table 3 depict that the accuracy of cattle counting based on full-appearance is 94% which is 4% higher than head. Therefore, the full-appearance detection outperforms the head detection approach in pasture situations. This discrepancy can be attributed to the difference of size and appearance of head and the whole body in cattle detection. The full-appearance detection for the feedlot also produces a good result with a counting accuracy of 92%, an AP of 91% for bounding-box prediction, 90% for mask prediction and a 0.95 F1 score. The standard deviation (SD) of APs for ten tests in each case is less than 0.01 and the counting numbers remain the same for ten tests in each case, indicating that the difference between the results is small so the results are stable and highly reliable.

4.3. Comparison with other state-of-the-art object detection algorithms

We compared the proposed Mask R-CNN model with three typical existing object detection methods: (1) Faster R-CNN, (2) Yolo v3, (3)

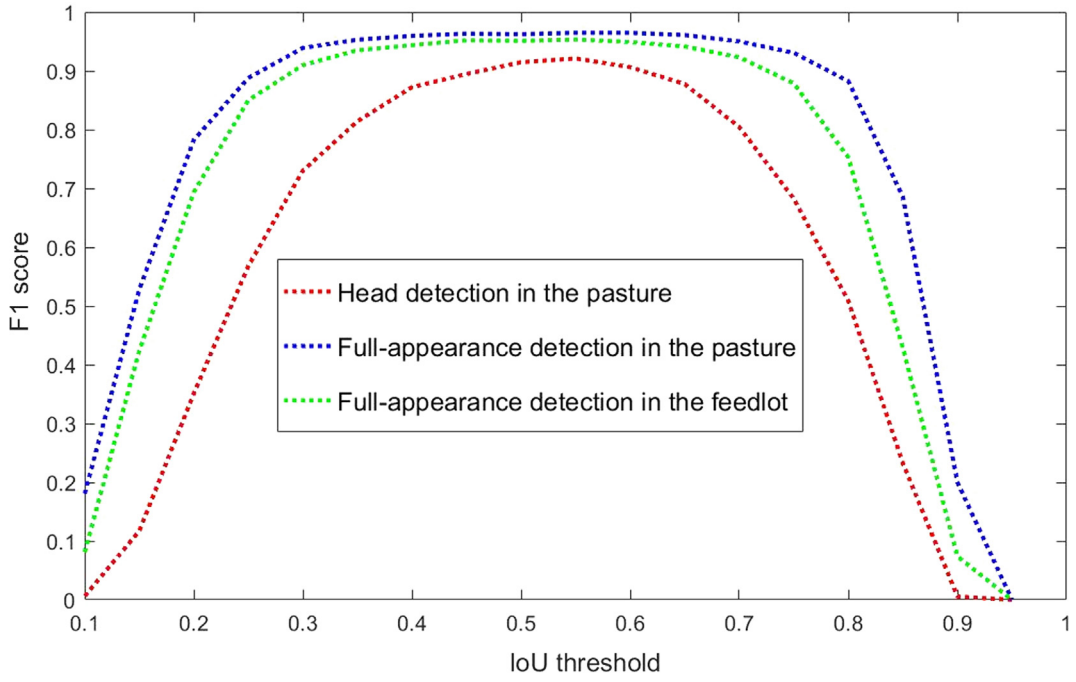

Fig. 5b. Comparisons of F1 scores over different IoU thresholds for three detection cases.

SSD. Among these competing methods, the first solution and Mask R-CNN model take the advantages of region proposals to improve the location accuracy, while the last two methods apply the regression-based technique. We evaluated the three competing methods on the same test images as Mask R-CNN collected from pastures and feedlot, and summarized the performances in term of counting accuracy and AP in Table 4. As we can see, Mask R-CNN used in this paper has achieved the best counting accuracy and AP in both two test datasets (marked in bold). The result indicates that Mask R-CNN is most effective in real-world datasets as the datasets are in different complex scenes with different density distributions and different degrees of occlusion. To facilitate the readers to visually observe the comparisons of results for different methods, we compared the predictions processed by the competing methods in Fig. 7. As mentioned in Section 2, the prediction for each cattle in the image using the Mask R-CNN presents both with bounding-box and mask, but three competing methods present just with bounding-box. Fig. 7(b) and (g) show the Mask R-CNN can detect the cattle precisely with occlusion, illumination and overlapping.

Table 2
AP and F1 scores of three detection cases.

Case	AP^{bb}	AP^{m}	F1	SD^{bb}	SD^{m}
Full-appearance detection (pasture)	0.95	0.94	0.96	0.008	0.006
Head detection (pasture)	0.86	0.84	0.91	0.007	0.008
Full-appearance detection (feedlot)	0.91	0.90	0.95	0.007	0.009

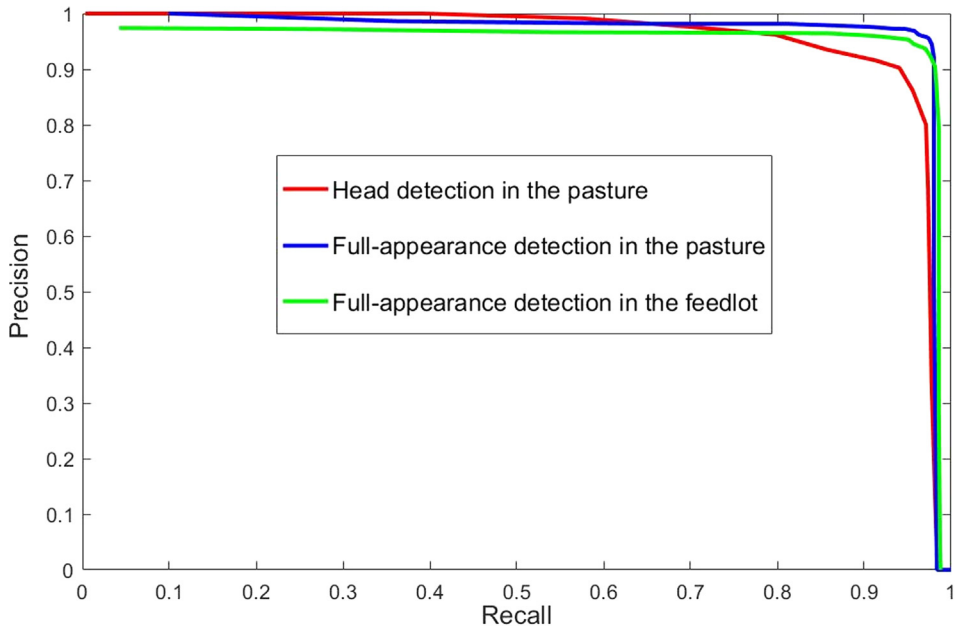

Fig. 6. Comparisons of precision-recall curves for three detection cases.

Table 3
Counting results of three detection cases.

Case	Ground truth	Detected	Counting error	Counting accuracy
Full-appearance detection (pasture)	857	912	0.06	0.94
Head detection (pasture)	857	952	0.10	0.90
Full-appearance detection (feedlot)	1064	1156	0.08	0.92

Table 4
Comparision of counting results with three competing methods.

Methods	Ground truth		Detected		Counting accuracy		AP	
	pasture	feedlot	pasture	feedlot	pasture	feedlot	pasture	feedlot
Faster R-CNN	857	1064	771	1013	0.90	0.95	0.89	0.8
Mask R-CNN	857	1064	912	1156	**0.94**	**0.92**	**0.96**	**0.94**
Yolo v3	857	1064	798	979	0.93	0.92	0.96	0.93
SSD	857	1064	725	883	0.85	0.83	0.92	0.85

5. Discussion

In this paper we evaluate Mask R-CNN, a state-of-the-art deep learning algorithm, for the detection and counting of cattle from quadcopter imagery. The key novelty of the work is the application of the Mask R-CNN algorithm and the demonstration of its effectiveness for this important livestock monitoring task. The essence of the detection in this paper is the binary classification with confidence and mask, that is, the result is cattle or not. Previous studies in cattle counting suffer the deviation of bounding-box and the challenge for mask detection (Rivas et al., 2018). A major advantage of the Mask R-CNN approach is the ability to perform both detection and instance segmentation of cattle within the imagery, this allows the development of further algorithms to perform tasks such as welfare monitoring from the imagery. Specifically, Mask R-CNN can also be used for key point detection (He et al., 2017), which can be used for real-time detection of behaviours of the animals to provide early warning for diseases like estrus (Dolecheck et al., 2015; Tian et al., 2013). Cattle instance segmentation in the paper is the first step towards real-time animal monitoring in farming environments that have different applications, such as early lameness detection (Viazzi et al., 2013) and other animal welfare improvements.

The Mask R-CNN detection performance was found to be affected by IoU threshold where the higher threshold lead to multiple predicted regions for one cow and the lower threshold resulted in lacking predicted region for other cattle. To evaluate the performance quantitatively and select the optimal threshold in this study, the F1 scores and Precision metrics were assessed over different thresholds. The results indicate that the threshold at 0.5 performs better all with average precision of more than 90%. Since there is no agreed standard threshold in object detection, this optimal threshold could be properly adjusted depending on the circumstances and applications in which it is used. For instance, IoU values above 0.70 for ecological camera trap data were considered well performing (Schneider et al., 2018) but IoU at 0.50 output may be better utilised for a person making an estimation in the wild, and both two using Faster R-CNN (Papandreou et al., 2017). However, current applications of Mask R-CNN including multi-person pose estimation (Chen et al., 2018) and beef cattle instance segmentation (Danish, 2018) achieve best results at around IoU = 0.5.

Considering the overlapping when many cattle gather together and repeat counting when some cattle are separated into several parts in different images, the full-appearance detection accuracy could be reduced to some extent. We learn from the successful practice of head detection of people to evaluate the effect of head detection for cattle. The comparisons of performance show that head detection results is unreliable where head detection methods achieve an accuracy of 90% for counting and full-appearance detection method achieves an accuracy of 94% for counting. The suggested main reason for this discrepancy of head detection performance is caused by multiple behaviours such as leaning over to graze or moving away from the drone in an opposite direction which then makes it difficult to detect the head. However, there is a possibility that head detection can be combined with facial features for individual cow identity.

In addition, we extended the full-appearance detection method to the feedlot case to evaluate the Mask R-CNN algorithms performance across various relevant cattle production scenarios. Also, we made the performance comparisons with other three competitive algorithms on the same datasets. The detection results presented illustrates that Mask R-CNN outperforms both in the counting accuracy and average precision, and the feedlot situation is a particularly challenging situation for the detection and counting of cattle even using Mask R-CNN and computer vision in general. Suggested reasons for this challenging situation include higher density of cattle and the similarity between the background pen surface colour and the cattle coat colour.

A wide variety of issues should also be considered including platform variation, sensor modality, costs and legal requirements associated with the monitoring of cattle using quadcopters like (Barbedo and Koenigkan, 2018). Barbedo & Koenigkan (2018) presented a strong case for the using Unmanned Aerial Systems (UAS) platforms for cattle detection and counting over extensive properties such as those in Brazil and Australia (Barbedo and Koenigkan, 2018). Whilst there is undoubtedly an important role for UAS monitoring of livestock on extensive properties, the outlook provided by Barbedo & Koenigkan (2018) downplayed the potential offered by small quad-copters already commercially available in many countries. Such technology has witnessed a strong interest from the livestock production industry. Quadcopters are used for a diverse range of tasks including mustering, checking infrastructure and stock welfare. Specific examples include the use of the DJI Phantom 4 quad-copter to monitor cattle over undulating and mountainous terrain in New South Wales, Australia, providing labour time-savings of 55 min (reported in Feedback magazine February/March 2017 https://www.mla.com.au/news-and-events/publications/feedback-magazine/2017-editions/) through monitoring and moving cattle in Nebraska, USA (https://www.agriculture.com/technology/livestock/up-in-the-air-cattle-management). Quad-copter drones are also of promise within feedlot operations where they can be used to monitor the number of stock at a facility (Condon, 2015). The high level of industry uptake is a strong indicator of the relevance of quad-copter technology for livestock monitoring.

6. Conclusion

Development of machine learning for object detection and instance segmentation is crucial to the vision system in a quadcopter. To

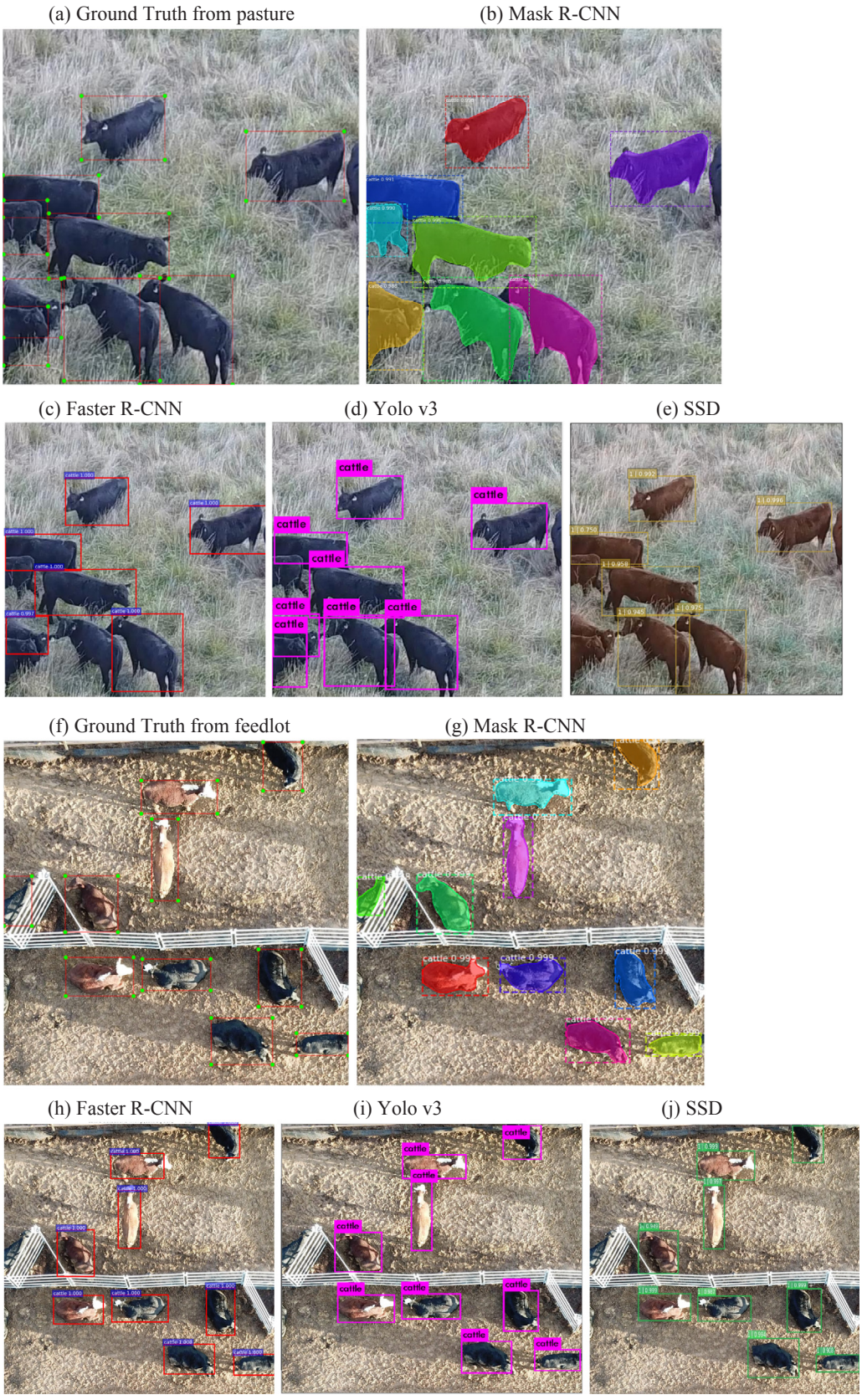

Fig. 7. Predictions on the test images for four object detection algorithms, images are from the pasture and feedlot. (Different colour of the mask and bounding-box have no special meanings).

establish the quadcopter machine vision system capable of monitoring livestock, we focused on cattle detection and counting which are important components of envisaged future technology. Performance of the Mask R-CNN model was assessed using manually annotated imagery acquired from a quadcopter and the compared metrics performed successfully across a range of relevant scenarios with average precision scores of 86%, 91%, 95% for bounding box and 84%, 90%, 94% for mask, and with a counting accuracy of 90%, 92%, 94, and a recall of 91%, 95%, 96%. The results presented indicate that Mask R-CNN could be utilised in practical settings as a method of livestock detection and counting using a quadcopter. Due to the high computational complexity of Mask R-CNN, the envisaged system would work best with a wireless link back to a central processing node either stationary or mobile to handle the higher computational requirements. The longer-term demonstration of Mask R-CNN paves the way for further algorithm innovations which could be utilised to process on the quadcopter.

From a practical precision livestock management perspective, this paper demonstrates that development of a key software component which could lead to quadcopters capable of autonomously identifying and quantifying livestock. Our research shows promising steps towards machine vision equipped livestock management quadcopters. In future work, we will concentrate on assessing Mask R-CNN performance over classification of livestock species and further explore the impact of stocking density on animal welfare.

CRediT authorship contribution statement

Beibei Xu: Methodology, Software, Investigation, Writing - original draft. **Wensheng Wang:** Conceptualization, Supervision, Resources, Funding acquisition, Writing - review & editing. **Greg Falzon:** Investigation, Supervision, Writing - review & editing. **Paul Kwan:** Investigation, Supervision, Writing - review & editing. **Leifeng Guo:** Data curation, Validation, Formal analysis. **Guipeng Chen:** Visualization, Software. **Amy Tait:** Resources, Writing - review & editing. **Derek Schneider:** Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This research was funded by Beijing Aokemei Technical Service Company Limited and also was supported by Fundamental Research Funds of Agricultural Information Institute of Chinese Academy of Agriculture Sciences, China (JBYW-AII-2019-19), General Project of Jiangxi Province Key Research and Development Plan (20192BBF60053) and Jiangxi Province Science Foundation for Youths (20192ACBL21023). Imagery of the feedlot animals was provided by a University of New England project funded by Meat and Livestock Australia (MLA) (University of New England Animal Ethics Approval Number AEC18-308) and we are grateful to three private farmlands in New England in Australia for their kindly support with data collection (University of New England Standard Operating Procedure W14 Camera Traps and Animal Ethics Approval Number AEC19-009).

References

Abd-Elrahman, A., Pearlstine, L., Percival, F., 2005. Development of pattern recognition algorithm for automatic bird detection from unmanned aerial vehicle imagery. Surv. Land Inform. Sci. 65, 37–45.

Andrew, W., 2017. Visual localisation and individual identification of holstein friesian cattle via deep learning. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2850–2859. https://doi.org/10.1109/ICCVW.2017.336.

Barbedo, Jayme Garcia Arnal, Koenigkan, Luciano Vieira, 2018. Perspectives on the use of unmanned aerial systems to monitor cattle. Outlook Agric. 47 (3), 214–222. https://doi.org/10.1177/0030727018781876.

Bengio, Y., Louradour, J., Collobert, R., Weston, J., 2009. Curriculum learning. In: Proceedings of the 26th Annual International Conference on Machine Learning, pp. 41–48. https://doi.org/10.1145/1553374.1553380.

Burić, M., Pobar, M., Ivašić-Kos, M., 2018. Object detection in sports videos. In: 2018 41st International Convention on Information and Communication Technology Electronics and Microelectronics (MIPRO), pp. 1034–1039.

Chabot, D., 2009. Systematic Evaluation of a Stock Unmanned Aerial Vehicle (UAV) System for Small-scale Wildlife Survey Applications. McGill University.

Chabot, D., Bird, D.M., 2015. Wildlife research and management methods in the 21st century: where do unmanned aircraft fit in? J. Unmanned Vehicle Syst. 3, 137–155. https://doi.org/10.1139/juvs-2015-0021.

Chabot, D., Francis, C.M., 2016. Computer-automated bird detection and counts in high-resolution aerial images: a review. J. Field Ornithol. 87, 343–359. https://doi.org/10.1111/jofo.12171.

Chamoso, P., Raveane, W., Parra, V., González, A., 2014. UAVs applied to the counting and monitoring of animals, Ambient Intelligence-Software and Applications. Springer, pp. 71–80. https://doi.org/10.1007/978-3-319-07596-9_8.

Chen, Y., Wang, Z., Peng, Y., Zhang, Z., Yu, G., Sun, J., 2018. Cascaded pyramid network for multi-person pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7103–7112.

Chrétien, L.-P., Théau, J., Ménard, P., 2015. Wildlife multispecies remote sensing using visible and thermal infrared imagery acquired from an unmanned aerial vehicle (UAV). Int. Arch. Photogramm. Rem. Sens. Spatial Inform. Sci. 40. https://doi.org/10.5194/isprsarchives-XL-1-W4-241-2015.

Condon, J., 2015. Drones Hold Promise in Cattle Applications, But Beware Some of the 'myths'. Beef Central.

Dalal, N., Triggs, B., 2005. Histograms of oriented gradients for human detection, Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on. IEEE, pp. 886–893. https://doi.org/10.1109/CVPR.2005.177.

Danish, M., 2018. Beef Cattle Instance Segmentation Using Mask R-Convolutional Neural Network. https://arxiv.org/pdf/1807.01972v2.pdf.

Descamps, S., Béchet, A., Descombes, X., Arnaud, A., Zerubia, J., 2011. An automatic counter for aerial images of aggregations of large birds. Bird Study 58, 302–308. https://doi.org/10.1080/00063657.2011.588195.

Dolecheck, K., Silvia, W., Heersche Jr, G., Chang, Y., Ray, D., Stone, A., Wadsworth, B., Bewley, J., 2015. Behavioral and physiological changes around estrus events identified using multiple automated monitoring technologies. J. Dairy Sci. 98, 8723–8731. https://doi.org/10.3168/jds.2015-9645.

Fan, R.-E., Lin, C.-J., 2007. A Study on Threshold Selection for Multi-label Classification. Department of Computer Science, National Taiwan University, pp. 1–23.

Frost, A., Schofield, C., Beaulah, S., Mottram, T., Lines, J., Wathes, C., 1997. A review of livestock monitoring and the need for integrated systems. Comput. Electron. Agric. 17, 139–159. https://doi.org/10.1016/s0168-1699(96)01301-4.

Gao, C., Li, P., Zhang, Y., Liu, J., Wang, L., 2016. People counting based on head detection combining Adaboost and CNN in crowded surveillance environment. Neurocomputing 208, 108–116. https://doi.org/10.1016/j.neucom.2016.01.097.

Girshick, R., 2015. Fast r-cnn. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1440–1448. https://doi.org/10.1109/ICCV.2015.169.

Girshick, R., Donahue, J., Darrell, T., 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, pp. 580–587. https://doi.org/10.1109/CVPR.2014.81.

Gonzalez, L.F., Montes, G.A., Puig, E., Johnson, S., Mengersen, K., Gaston, K.J., 2016. Unmanned Aerial Vehicles (UAVs) and artificial intelligence revolutionizing wildlife monitoring and conservation. Sensors 16, 97. https://doi.org/10.3390/s16010097.

Grenzdörffer, G., 2013. UAS-based automatic bird count of a common gull colony. Int. Arch. Photogramm. Rem. Sens. Spatial Inform. Sci. 1, W2. https://doi.org/10.5194/isprsarchives-XL-1-W2-169-2013.

Handcock, R.N., Swain, D.L., Bishop-Hurley, G.J., Patison, K.P., Wark, T., Valencia, P., Corke, P., O'Neill, C.J., 2009. Monitoring animal behaviour and environmental interactions using wireless sensor networks, GPS collars and satellite remote sensing. Sensors 9, 3586–3603. https://doi.org/10.3390/s90503586.

He, Kaiming, Gkioxari, Georgia, Dollar, Piotr, Girshick, Ross, 2017. Mask R-CNN. IEEE Trans. Pattern Anal. Mach. Intell. 42 (2), 386–397. https://doi.org/10.1109/TPAMI.3410.1109/TPAMI.2018.2844175.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778. https://doi.org/10.1109/CVPR.2016.90.

Hodgson, A., Kelly, N., Peel, D., 2013. Unmanned aerial vehicles (UAVs) for surveying marine fauna: a dugong case study. PLoS One 8, e79556. https://doi.org/10.1371/journal.pone.0079556.

Hollings, T., Burgman, M., van Andel, M., Gilbert, M., Robinson, T., Robinson, A., 2018. How do you find the green sheep? A critical review of the use of remotely sensed imagery to detect and count animals. Methods Ecol. Evol. 9, 881–892. https://doi.org/10.1111/2041-210X.12973.

Huang, J., Rathod, V., Sun, C., Zhu, M., Korattikara, A., Fathi, A., Fischer, I., Wojna, Z., Song, Y., Guadarrama, S., 2017. Speed/accuracy trade-offs for modern convolutional object detectors. IEEE CVPR. https://doi.org/10.1109/CVPR.2017.351.

Jail, S., Chinawalkar, M., Khedkar, R., Rathi, Y., Vidap, P., 2018. Quality assessment of mangoes using computer vision and machine learning. Int. J. Eng. Comput. Sci. 7, 23908–23913.

Kellenberger, B., Marcos, D., Tuia, D., 2018. Detecting mammals in UAV images: best practices to address a substantially imbalanced dataset with deep learning. Rem. Sens. Environ. 216, 139–153. https://doi.org/10.1016/j.rse.2018.06.028.

Koski, W.R., Allen, T., Ireland, D., Buck, G., Smith, P.R., Macrander, A.M., Halick, M.A., Rushing, C., Sliwa, D.J., McDonald, T.L., 2009. Evaluation of an unmanned airborne system for monitoring marine mammals. Aquatic Mammals 35, 347. https://doi.org/10.1578/am.35.3.2009.347.

Lee, A., 2015. Comparing Deep Neural Networks and Traditional Vision Algorithms in Mobile Robotics. Swarthmore University.

Lhoest, S., Linchant, J., Quevauvillers, S., Vermeulen, C., Lejeune, P., 2015. HOW MANY HIPPOS (HOMHIP): algorithm for automatic counts of animals with infra-red thermal imagery from UAV. International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences 40. https://doi.org/10.5194/isprsarchives-XL-3-W3-355-2015.

Li, Z., Peng, C., Yu, G., Zhang, X., Deng, Y., Sun, J., 2017. Light-head r-cnn: In defense of two-stage object detector. arXiv preprint arXiv:1711.07264.

Liaghat, S., Balasundram, S.K., 2010. A review: the role of remote sensing in precision agriculture. Am. J. Agric. Biol. Sci. 5, 50–55. https://doi.org/10.3844/ajabssp.2010.50.55.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., Berg, A.C., 2016. Ssd: Single shot multibox detector, European Conference on Computer Vision. Springer, pp. 21–37. https://doi.org/10.1007/978-3-319-46448-0_2.

Longmore, S., Collins, R., Pfeifer, S., Fox, S., Mulero-Pázmány, M., Bezombes, F., Goodwin, A., De Juan Ovelar, M., Knapen, J., Wich, S., 2017. Adapting astronomical source detection software to help detect animals in thermal images obtained by unmanned aerial systems. Int. J. Rem. Sens. 38, 2623–2638. https://doi.org/10.1080/01431161.2017.1280639.

Marsh, J.R., Gates, R.S., Day, G.B., Aiken, G.E., Wilkerson, E.G., 2008. Assessment of an injectable RFID temperature sensor for indication of horse well-being, 2008 Providence, Rhode Island, June 29–July 2, 2008. American Society of Agricultural and Biological Engineers, p. 1. https://doi.org/10.13031/2013.24845.

Mejias, L., Duclos, G., Hodgson, A., Maire, F., 2013. Automated marine mammal detection from aerial imagery, Oceans-San Diego, 2013. IEEE, pp. 1–5. https://doi.org/10.1016/j.oceaneng.2012.11.007.

Neethirajan, S., 2017. Recent advances in wearable sensors for animal health management. Sens. Bio-Sens. Res. 12, 15–29. https://doi.org/10.1016/j.sbsr.2016.11.004.

Neethirajan, S., Tuteja, S.K., Huang, S.-T., Kelton, D., 2017. Recent advancement in biosensors technology for animal and livestock health management. Biosens. Bioelectron. 98, 398–407. https://doi.org/10.1016/j.bios.2017.07.015.

Norouzzadeh, M.S., Nguyen, A., Kosmala, M., Swanson, A., Palmer, M.S., Packer, C., Clune, J., 2018. Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning. Proc. Natl. Acad. Sci. 201719367. https://doi.org/10.1073/pnas.1719367115.

Papandreou, G., Zhu, T., Kanazawa, N., Toshev, A., Tompson, J., Bregler, C., Murphy, K., 2017. Towards accurate multi-person pose estimation in the wild, CVPR, p. 6. https://doi.org/10.1109/CVPR.2017.395.

Pathare, S.P., 2015. Detection of Black-backed Jackal in Still Images. Stellenbosch University, Stellenbosch.

Qiao, Y., Truman, M., Sukkarieh, S., 2019. Cattle segmentation and contour extraction based on Mask R-CNN for precision livestock farming. Comput. Electron. Agric. 165, 104958.

Radovic, M., Adarkwa, O., Wang, Q., 2017. Object recognition in aerial images using convolutional neural networks. J. Imag. 3, 21. https://doi.org/10.3390/jimaging3020021.

Redmon, J., Divvala, S., Girshick, R., Farhadi, A., 2016. You only look once: Unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 779–788. https://doi.org/10.1109/CVPR.2016.91.

Redmon, J., Farhadi, A., 2018. YOLOv3: An Incremental Improvement, arXiv e-prints. https://arxiv.org/pdf/1804.02767.pdf.

Ren, S., He, K., Girshick, R., Sun, J., 2017. Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans. Pattern Anal. Mach. Intell. 39 (6), 1137–1149. https://doi.org/10.1109/TPAMI.2016.2577031.

Rey, N., Volpi, M., Joost, S., Tuia, D., 2017. Detecting animals in African Savanna with UAVs and the crowds. Remote Sens. Environ. 200, 341–351. https://doi.org/10.1016/j.rse.2017.08.026.

Rivas, A., Chamoso, P., González-Briones, A., Corchado, J., 2018. Detection of cattle using drones and convolutional neural networks. Sensors 18, 2048. https://doi.org/10.3390/s18072048.

Ruiz-Garcia, L., Lunadei, L., Barreiro, P., Robla, I., 2009. A review of wireless sensor technologies and applications in agriculture and food industry: state of the art and current trends. Sensors 9, 4728–4750. https://doi.org/10.3390/s90604728.

Russell, B.C., Torralba, A., Murphy, K.P., Freeman, W.T., 2008. LabelMe: a database and web-based tool for image annotation. Int. J. Comput. Vision 77, 157–173. https://doi.org/10.1007/s11263-007-0090-8.

Sa, I., Ge, Z., Dayoub, F., Upcroft, B., Perez, T., McCool, C., 2016. Deepfruits: a fruit detection system using deep neural networks. Sensors 16, 1222. https://doi.org/10.3390/s16081222.

Sadgrove, E.J., Falzon, G., Miron, D., Lamb, D., 2017. Fast object detection in pastoral landscapes using a colour feature extreme learning machine. Comput. Electron. Agric. 139, 204–212. https://doi.org/10.1016/j.compag.2017.05.017.

Sadgrove, E.J., Falzon, G., Miron, D., Lamb, D.W., 2018. Real-time object detection in agricultural/remote environments using the multiple-expert colour feature extreme learning machine (MEC-ELM). Comput. Ind. 98, 183–191. https://doi.org/10.1016/j.compind.2018.03.014.

Schneider, S., Taylor, G.W., Kremer, S.C., 2018. Deep Learning Object Detection Methods for Ecological Camera Trap Data. arXiv preprint arXiv:1803.10842.

Sellier, N., Guettier, E., Staub, C., 2014. A review of methods to measure animal body temperature in precision farming. Am J. Agric. Sci. Technol. 2, 74–99. https://doi.org/10.7726/ajast.2014.1008.

Sommer, L., Schumann, A., Schuchert, T., Beyerer, J., 2018. Multi feature deconvolutional faster R-CNN for precise vehicle detection in aerial imagery. In: IEEE Winter Conference on Applications of Computer Vision, https://doi.org/10.1109/WACV.2018.00075.

Stein, M., Bargoti, S., Underwood, J., 2016. Image based mango fruit detection, localisation and yield estimation using multiple view geometry. Sensors 16, 1915. https://doi.org/10.3390/s16111915.

Tian, F.Y., Wang, R.R., Liu, M.C., Wang, Z., Li, F.D., Wang, Z.H., 2013. Oestrus detection and prediction in dairy cows based on neural networks. Trans. CSAM 44, 5. https://doi.org/10.6041/j.issn.1000-1298.2013.S1.050.

Van Nuffel, A., Zwertvaegher, I., Van Weyenberg, S., Pastell, M., Thorup, V.M., Bahr, C., Sonck, B., Saeys, W., 2015. Lameness detection in dairy cows: Part 2. Use of sensors to automatically register changes in locomotion or behavior. Animals 5, 861–885. https://doi.org/10.3390/ani5030387.

Vermeulen, C., Lejeune, P., Lisein, J., Sawadogo, P., Bouché, P., 2013. Unmanned aerial survey of elephants. PLoS One 8, e54700. https://doi.org/10.4236/tel.2017.72015.

Viazzi, S., Bahr, C., Schlageter-Tello, A., Van Hertem, T., Romanini, C., Pluk, A., Halachmi, I., Lokhorst, C., Berckmans, D., 2013. Analysis of individual classification of lameness using automatic measurement of back posture in dairy cattle. J. Dairy Sci. 96, 257–266. https://doi.org/10.3168/jds.2012-5806.

Yu, X., Wang, J., Kays, R., Jansen, P.A., Wang, T., Huang, T., 2013. Automated identification of animal species in camera trap images. EURASIP J. Image Video Process. 2013, 52. https://doi.org/10.1186/1687-5281-2013-52.

Zhang, W., 1988. Shift-invariant pattern recognition neural network and its optical architecture. Proceedings of Annual Conference of the Japan Society of Applied Physics.

Zhang, W., Itoh, K., Tanida, J., Ichioka, Y., 1990. Parallel distributed processing model with local space-invariant interconnections and its optical architecture. Appl. Opt. 29, 4790–4797. https://doi.org/10.1364/AO.29.004790.

Zhou, D., 2014. Real-time Animal Detection System for Intelligent Vehicles. University of Ottawa.