Computers and Electronics in Agriculture 165 (2019) 104958

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

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

Cattle segmentation and contour extraction based on Mask R-CNN for precision livestock farming

Yongliang Qiao*, Matthew Truman, Salah Sukkarieh

Australian Centre for Field Robotics (ACFR), Faculty of Engineering, The University of Sydney, NSW 2006, Australia

ARTICLE INFO

Keywords:
Precision livestock farming
Instance segmentation
Cattle contour
Deep learning
Mask R-CNN

ABSTRACT

In precision livestock farming, computer vision based approaches have been widely used to obtain individual cattle health and welfare information such as body condition score, live weight, activity behaviours. For this, precisely segmenting each cattle image from its background is a prerequisite, which is an important step towards obtaining real-time individual cattle information. In this paper, an instance segmentation approach based on a Mask R-CNN deep learning framework is proposed to solve cattle instance segmentation and contour extraction problems in a real feedlot environment. The proposed approach consists of the following steps: key frame extraction (detect the huge cattle motion frames), image enhancement (reduce the illumination and shadow influence), cattle segmentation and body contour extraction. We trained and tested the proposed approach on a challenging cattle image dataset. According to the experimental results, the proposed approach can render fairly desirable cattle segmentation performance with 0.92 Mean Pixel Accuracy (MPA) and achieve contour extraction with an Average Distance Error (ADE) of 33.56 pixel, which is better than that of the state-of-the-art SharpMask and DeepMask instance segmentation methods.

1. Introduction

With the increasing consumption demand on livestock production, global livestock industry has to feed more animals with limited environmental resources (e.g. farmland, facilities) and the shortage of livestock labor-force (Thornton, 2010; Van Hertem et al., 2017). In this situation, precision livestock farming plays an important role in achieving high efficiency with low cost, in an environmentally sustainable way (Hariharan et al., 2014; Halachmi and Guarino, 2016). In doing so, obtaining welfare, wellbeing and behaviour information of individual cattle makes a significant contribution in livestock farming management decision making (He et al., 2016; Bos et al., 2018). Nowadays, cameras as a type of non-contact and cheap sensors, are widely used to monitor cattle (Zin et al., 2016). In vision based approaches, image segmentation is a prerequisite. After each individual animal' contour is extracted from the background, the subsequent image analysis technologies enable farmers to monitor variables (i.e. length, width and back area) relevant to the health, welfare and productivity of individual animal throughout their life cycle, and design farming management strategies in more efficient ways (Schoder and Staunfenbiel, 2006).

From the segmented images, a variety of studies have been conducted to extract visual features to perform animal welfare evaluation and behaviour analysis such as body measurement (Gomes et al., 2016; Zhang et al., 2018), body condition score estimation (Hansen et al., 2018; Lynn et al., 2017), live weight prediction (Stajnko et al., 2008; Tebug et al., 2018) and lameness detection (Nilsson et al., 2015; Nasirahmadi et al., 2017). Here, extraction of all visual-related features (body length, width, curvature, posture and so on) relies heavily on the segmented images. It is obvious that the accuracy and efficiency of image segmentation play a significant role in the further image analysis in vision based real-time automatic individual cattle monitoring and welfare evaluation. However, traditional frame difference (Zhan et al., 2007) or optical flow (Zitnick et al., 2005) based approaches struggle to achieve high segment accuracy under complex outdoor environments. The changing illumination conditions, shadows and dynamic background (e.g. walking farm workers, dirty ground, other animals and cattle trucks) cause difficulties which will affect the quality of cattle image segmentation (Ter-Sarkisov et al., 2018).

In recent years, deep convolutional neural networks (CNNs) with strong feature learning abilities have been widely used in the field of computer vision (Qiao et al., 2019; Kumar et al., 2018). Thanks to the training on massive annotated visual data, those features learned by CNNs carry rich spatial and semantic information which can be used in image segmentation (Pinheiro et al., 2015; Andrew et al., 2017). Li et al. (2016) proposed an iterative instance segmentation that is able to

* Corresponding author.
E-mail address: yongliang.qiao@sydney.edu.au (Y. Qiao).

https://doi.org/10.1016/j.compag.2019.104958
Received 4 March 2019; Received in revised form 2 August 2019; Accepted 12 August 2019
Available online 21 August 2019

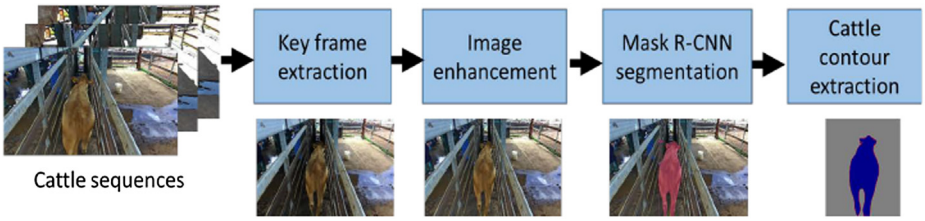


Fig. 1. Flowchart of the proposed approach for cattle segmentation and contour extraction.

learn implicit shape priors to improve the predicted pixel-wise labeling quality. In addition, DeepMask (Pinheiro et al., 2015) and SharpMask (Pinheiro et al., 2016) generated segmentation object proposals directly from image pixels and then classified them into different categories. More recently, He et al. (2017) proposed a Mask R-CNN framework for keypoint detection and instance segmentation. All these recent research achievements show the viability of CNNs based approaches for cattle segmentation under complex feedlot environments.

In this paper, in order to achieve high accuracy segmentation and contour extraction performance in complex background environments, an automatic cattle segmentation and contour extraction method based on Mask R-CNN is proposed. As demonstrated in Fig. 1, the whole proposed approach consists of several steps: firstly, key frames are extracted from surveillance videos; secondly, an enhancement method is applied to improve the image quality; finally cattle segmentation and contour extraction are conducted. This work is a step towards real-time cattle welfare evaluation in precision livestock farming applications such as body condition scoring (Krukowski, 2009), lameness detection (Gardenier et al., 2018), animal live weight estimation and reduction of cattle management costs (Song et al., 2018; Kashiha et al., 2014). The paper is organized as follows: Section 2 briefly reviews some detection and segmentation related deep learning networks, Section 3 introduces the dataset acquisition system and proposed approach, Section 4 illustrates the performance evaluation method, Section 5 presents all experimental results, and conclusions and suggestions for future research are given in Section 6.

2. Related work

Algorithms based on the deep learning networks have led to dramatic advances in the field of object detection (Wang et al., 2018; Norouzzadeh et al., 2018), semantic segmentation (Long et al., 2015) and instance segmentation (Pinheiro et al., 2015; He et al., 2017).

Most object detection methods generate a bounding box for each detect target and then classify these objects (Ren et al., 2015). The Region-based Convolutional Neural Network (R-CNN) method generates region proposals by selective search and then classifies object proposals using a deep CNN (Girshick et al., 2015). However, R-CNN extraction of the proposal region features are costly. In Faster R-CNN (Ren et al., 2015), region proposals are generated by a separate Region Proposal Network (RPN). It takes an image as input and outputs a set of rectangular object proposals, each with an abjectness score (Ren et al., 2015). Faster R-CNN is composed of two branches: the first branch is a RPN that proposes candidate object bounding boxes, while the second branch is the Fast R-CNN detector, which extracts features from each candidate box and performs classification and bounding-box regression (Girshick, 2015).

Development from the object detection, image segmentation tries to delineate classes of objects at pixelwise level. There are two different categories: semantic and instance segmentation. Semantic segmentation is a pixel-labeling task which labels each pixel of an image for specific classes of objects; it does not distinguish the objects from the same class (Li et al., 2016). A widely used semantic segmentation model is Fully Convolutional Networks (FCN); it is a CNN variant which transforms image pixels to pixel categories. Instance segmentation identifies each object instance using a mask representation in the image, it realizes the object class prediction and mask extraction at the same time (Zhao et al., 2018). Instance segmentation usually consists of three steps: (a) proposal regions identification using RPN, where RPN is a fully convolutional network that simultaneously predicts object bounds and objectness scores at each position, (b) object class prediction, and (c) its mask extraction. Here, the mask extraction encodes an input object's spatial layout through the convolutional operation. In recent years, some instance segmentation methods have been proposed (Chen et al., 2018). Bai and Urtasun (2017) combined watershed transform and modern deep learning to produce an energy map and realized object instances segmentation by a single energy level cutting. Ter-Sarkisov et al. (2018) extended a fully convolutional network to realize beef cattle segmentation.

3. Material and methods

In this section, the cattle image acquisition platform, experimental dataset and overview of the proposed approach are introduced.

3.1. Data acquisition

In this study, beef cattle in Australian Country Choice's Brisbane Valley feedlot were our research target. All cattle are kept in some large outdoor pens and each of them has an ear tag for identification purposes. They are walking and laying freely on the ground in the open-air pens, if necessary they will be driven to the crush station for live weighing, blood sampling or ear tag installation. The whole data acquisition platform, as shown in Fig. 2, was placed nearby the cattle crush. It mainly consists of two stereo ZED cameras (the camera horizontal field of view is 110°), a GPU-equipped small PC and a high-speed volume data storage disk. In our experiment, only the left image of the rear view ZED camera was used; the image resolution was set to 1920 × 1080 with 30fps frame rate. During data collection, when beef cattle were driven into the crush, they were easily frightened and tried to run away from the crush due to the fear and stress. Therefore, a high frame acquisition rate (30fps) was adopted to reduce the influences of motion blur.

In our project, cattle data were collected in three different time (induction, middle point and end point) on 20 March, 30 April and 30 May respectively in 2018. Image sequences were acquired when the

Fig. 2. Experimental setup at Brisbane Valley feedlot showing rear view camera (clamped to overhead bar) and side view camera (on tripod). PC is Neousys Nuvo 6108GC.

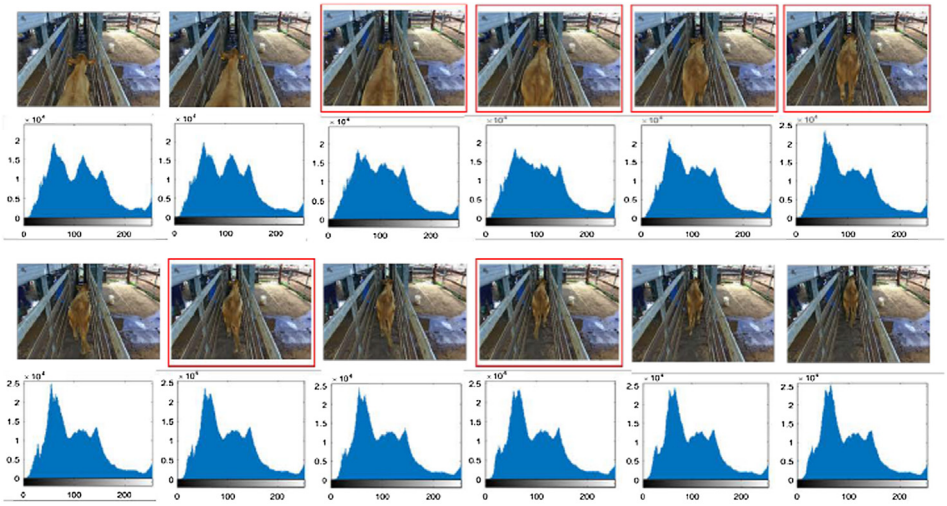


Fig. 3. Examples of the selected key frames (red rectangular boxed images). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

cattle were walking on the pathway (from right to left in Fig. 2. Cattle live weight varied from 330 kg to 550 kg as body size and mass changed in these three months. Standard deviation of cattle live weight in the training dataset is around 45 kg. This cattle dataset is really challenging for the image segmenter when considering the following aspects: (1) The animals change posture frequently when they are driven into the cattle crush; (2) High similarity between animals from same species due to the similar coat color; (3) Lighting is changing with time of day. Images acquired in the morning are low-light, while they are over-illuminated and have strong shadows at noon. These lighting problems can make machine learning algorithms erroneously learn these patches or shadows as animals' features. (4) The complex background (including the cattle crush and ground) is difficult to be distinguished from the segmented cattle. In some cases, the segmented animal instance nearly blends with the soil background.

3.2. Overview of the proposed approach

For individual cattle monitoring and their welfare measurement in vision based precision livestock farming, effective cattle segmentation is the prerequisite for further image analysis (Oczak et al., 2013). The Mask R-CNN based approach is proposed to realize cattle instance segmentation in the complex feedlot environment.

As illustrated in Fig. 1, for a given cattle image sequence, key frames and normal frames are firstly determined through the key frame extraction step. In order to reduce the influence of illumination and shadows, image enhancement is then applied to the selected key frames. After this, the Mask R-CNN model is constructed and trained to detect and segment the cattle body area. These segmentation results could be used in some follow-up works such as obtaining further cattle body parameters (e.g. length, width and back area). In the end, based on the segmented cattle body areas, cattle contour lines are extracted. The details of each step are illustrated below.

3.3. Key frame extraction from image sequences

In the image sequences, cattle were not walking or moving in some moments due to watch or hoof diseases. Thus the collected dataset contains many repeated images. On the other hand, farmers are more interested in the frames which contain cattle motion or posture changing information for the welfare evaluation and behaviour analysis (Tscharke and Banhazi, 2016). If the same method is applied to each frame, the whole system efficiency will be low. In order to reduce the repeated images, focus on the key motion information and improve the system efficiency, each frame in the image sequences is classified into key frame F_{key} or normal frame F_{normal}. Key frames, are the images where the cattle posture or behaviour exhibits changes, usually contain more motion information. These key frames are important indicators for cattle behaviour analysis or health evaluation (Tscharke and Banhazi, 2016). Absolute histogram difference based approach (Sheena and Narayanan, 2015) is a simple and effective way to conduct key frame extraction. As the cattle are moving, the histogram of each image frame is varying. According to the histogram difference of cattle movements, key frames can be selected. The algorithm of key frame selection is as follows:

Algorithm 1. Key frame extraction in image sequence.

Inputs:
F_i {The i-th image};
N {Image number of the sequence};
Outputs:
F_{key}, F_{normal} {key frame and normal frame}; bf Algorithm:
for $i \leftarrow$ 1 to N **do**
H_i, $H_{i+1} \leftarrow$ Image histogram computation;
$D_{i,i+1} \leftarrow$ Absolute histogram difference between two consecutive frames;
if $D_{i,i+1} > Th$;F_i is marked as a key frame F_{key}.
if $D_{i,i+1} \leqslant Th$;F_i is marked as a normal frame F_{normal}.
end for

Examples of the selected key frames and their corresponding histograms can be seen in Fig. 3. We set a threshold Th of 50 between the histograms of two consecutive image frames; thus the repeated images are removed and the frames containing cattle posture and motion information are selected as the key frames.

3.4. Image enhancement

Since cattle image collection usually lasts a long time (one day for each time data collection), the outdoor non-uniform illumination and shadows have a large influence on the image quality. According to the Retinex theory (Land, 1986), image is a product of two components such as the illumination image (low-frequency characteristics) and the reflection image (high-frequency detail information such as edge and texture). Therefore, in order to avoid huge differences in brightness between image frames, a two-dimensional gamma function-based image adaptive correction algorithm (Liu et al., 2016) is used to remove non-uniform illumination and shadow influence to improve image

quality. The main image enhancement steps are as follows.

- Illumination components extraction: We firstly transfer the RGB color space into HSV space, thus the operation of brightness V channel does not affect the image color and texture information. Then multi-scale Gaussian function $G(x, y)$ is used to extract illumination components of the scene in brightness V channel:

$$\begin{cases} I(x, y) = \sum_{i=1}^{N} \omega_i [F(x, y) G(x, y)] \\ G(x, y) = \lambda \exp(-\frac{x^2 + y^2}{c^2}) \end{cases} \quad (1)$$

  where $F(x, y)$ is the input image, $I(x, y)$ is illumination component values that extracted and weighted by multi-scale Gaussian functions $G(x, y)$ at the point (x, y), ω_i is the weight coefficient of the ith scale Gaussian function, N is the number of scales, c is the scale factor and λ is the normalization constant. In our study, by balancing the precision and computation costs, three different scales factors c (15, 80 and 250) with 1/3 weight coefficients respectively are used in this study.
- Adaptive illumination adjustment: After extracting the illumination component of the scene, unevenness correction function can be constructed according to the distribution characteristics of the illumination component. In this study, a two-dimensional gamma function $O(x, y)$ is constructed to reduce the luminance value of the over-illuminated region, and increase the luminance value of the over-dark region image:

$$O(x, y) = 255(\frac{F(x, y)}{255})^r, \quad r = 0.5^{\frac{m - I(x,y)}{m}} \quad (2)$$

  where $O(x, y)$ is the brightness value of the corrected output image; r is the index value for brightness enhancement, which contains the illumination component characteristics of the image; m is the brightness mean of the illumination component.
- RGB image reconstruction: Based on the new brightness value $O(x, y)$, the enhanced image can be obtained by reconstructing the RGB image from the HSV color space. In the enhanced image, the influence of uneven illumination is reduced and the quality is improved.

The performance of image enhancement is displayed in Fig. 4. By comparing the enhanced images with the raw images, it can be seen that the details of the low-light and over-illuminated region of cattle images are significantly improved. In addition, clarity of shadows covering cattle images is also improved due to effective brightness balance of 2D gamma function.

3.5. Mask R-CNN based cattle segmentation

Our study focus is on segmenting of individual cattle image from background, which is a typical instance segmentation task. The proposed Mask R-CNN based cattle segmentation is performed by combining detection (individual cattle classify and localization) and semantic segmentation (identifying cattle pixels) together.

3.5.1. Mask R-CNN network

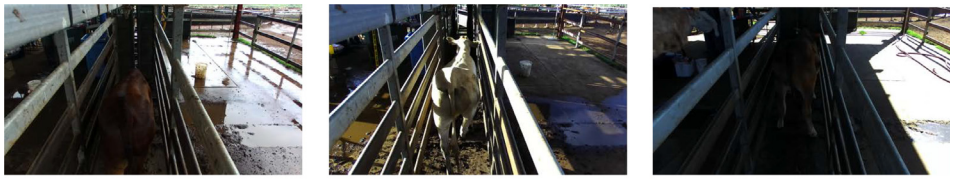
The Mask R-CNN (He et al., 2017), as a flexible instance segmentation model, is improved from the Faster R-CNN (Girshick, 2015) by adding a segmentation mask generating branch. The framework of Mask R-CNN based cattle segmentation is illustrated in Fig. 5. It has two parts: (1) Convolutional backbone part: the convolutional backbone is responsible for feature extraction over an entire image; (2) Head part: it performs bounding-box recognition (classification and regression) and mask prediction. The RPN network computes the region proposals and then RoIAlign (He et al., 2017) extracts features from each proposal and performs two parallel operations. One gets the cattle detection, classification and bounding box regression after FC layers. The other after RoIAlign outputs high accuracy segmentation masks.

Since Feature Pyramid Network (FPN) takes advantage of both high-resolution feature maps and high -level semantic information for accurate localization (Lin et al., 2017), in our study, a FPN based ResNet-101 (Hariharan et al., 2014) network is used as a backbone to achieve gains in both accuracy and speed (Lin et al., 2017; Hariharan et al., 2014), while Faster R-CNN with ResNet acts as the head architecture.

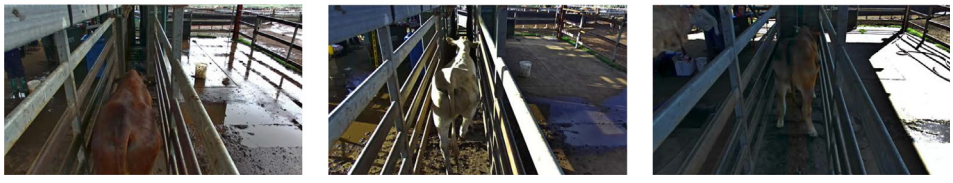
For each cattle image, CNN features are firstly extracted using ResNet-101, then RPN uses a sliding window approach (Girshick, 2015) on the feature maps to compute the bounding box proposals. Before the next step, RoIAlign are used to map arbitrarily sized spatial regions of interest in the features to a fixed spatial resolution by using bilinear interpolation. Finally, Mask R-CNN head part predicts object class, refines the bounding box localization and generates segmentation masks simultaneously.

3.5.2. Muti-feature exaction

In the Mask R-CNN based approach, the performance of object (cattle) detection and segmentation can be improved if more feature information is used. Therefore, the powerful ResNet101 (Hariharan et al., 2014) was selected as backbone. It consists of 5 stages, corresponding to 5 different scales of feature map [$C1$, $C2$, $C3$, $C4$, $C5$]. These feature maps are used to establish the feature pyramid of FPN network, and get new features [$P1$, $P2$, $P3$, $P4$, $P5$] respectively. In real experiments, $P1$ is not used considering it takes huge time to calculate the feature map corresponding to $C1$. Actually, through downsampling $P5$, $P6$ is obtained and it is used to replace $P1$. The specific



(a) Raw images



(b) Ehanced images

Fig. 4. Comparison of the raw and enhanced images. The enhanced images' quality are improved by removing the non-uniform illumination effects.

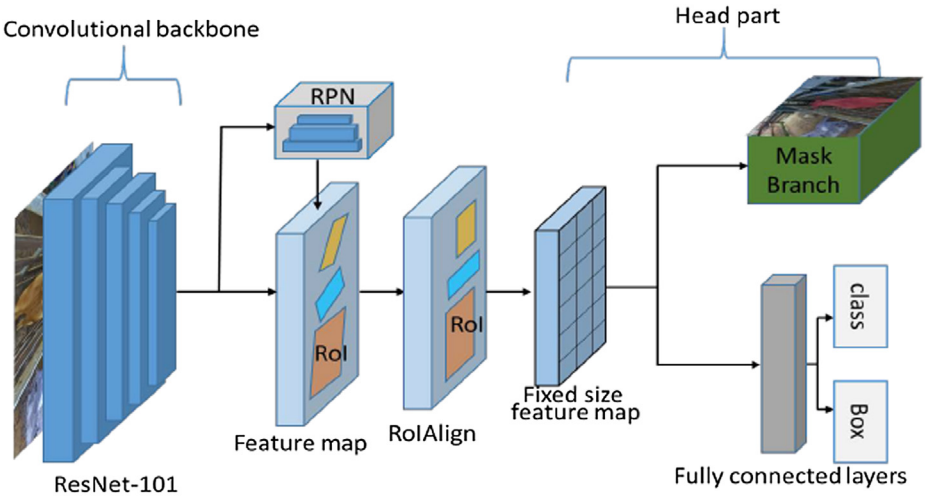


Fig. 5. Framework of Mask R-CNN based cattle segmentation.

correspondence of feature map is shown as follows:

$$\begin{cases} P_2 = conv(sum(upsample(P_3, conv(C_2)))) \\ P_3 = conv(sum(upsample(P_4, conv(C_3)))) \\ P_4 = conv(sum(upsample(P_5, conv(C_4)))) \\ P_5 = conv(conv(C_5)) \\ P_6 = downsample(P_5) \end{cases} \tag{3}$$

where *conv* represents the convolution operation, *sum* represents the element-by-element alignment operation, *upsample* and *downsample* represent upsampling and downsampling operation respectively.

3.5.3. Proposed region generation and RoIAlign operation

The obtained feature maps [*P*2, *P*3, *P*4, *P*5, *P*6] are sent to the RPN, which uses a sliding window to scan the feature maps and find the RoI areas (regions) where the cattle exist. Each obtained RoI area is a rectangle (Anchor) on the image. The RPN network outputs the category of Anchor and the correction data for the border coordinates. The first output is used to judge whether the Anchor is a foreground or background. If it is a foreground, it means that there is an object (cattle) in the Anchor box, although the object (cattle) may not be perfectly at the center of the box. Therefore, correction percentages, the second RPN network output, is used to correct the bounding box to better fit the object (cattle).

After the RPN network processing and prediction, a series of bounding boxes can be obtained and their position and size are corrected. If there are multiple bounding boxes overlap each other, the non-maximum suppression (NMS) is applied to get the bounding box with a higher foreground score and pass it to the next stage.

Before the next stage, in order to produce a standard-sized output for input to a classifier, RoIAlign (He et al., 2017) is used to adjust the size of the Anchor box to a fixed size. It aligns the extracted features with the original region proposal properly, and helps to produce better pixel segmentation results. In the back propagation of the RoIAlign layer, $i^*(r, j)$ is the coordinate position of a floating point number (sample points calculated during forward propagation), in the feature map before pooling, each point with the horizontal and vertical coordinates of $i^*(r, j)$ are less than 1 should accept the gradient of the corresponding point $y_{r,j}$, and the back propagation formula of the RoIAlign layer is as follows:

$$\begin{cases} \frac{\partial L}{\partial x_i} = \sum_r \sum_j [d(i, i^*(r, j)) < 1](1 - \triangle h)(1 - \triangle w)\frac{\partial L}{\partial y_{r,j}} \end{cases} \tag{4}$$

where $d(\cdot)$ represents the distance between two points, $\triangle h$ and $\triangle w$ represent the difference between x_i and $x_{i*}(r, j)$. Through the RoIAlign process, the extracted features with the input image are aligned properly, the subpixel misalignment problem when defining corresponding region between the region proposal and the feature map is solved, which makes pixel segmentation more accurate.

3.5.4. Cattle instance segmentation and loss function

The features obtained by RoIAlign are fed to the Full Connected (FC) layer for classification and bounding box regression, and also fed to the convolution layer for segmentation. The classification is done by passing the output from the FC layer that uses all the features through the softmax layer. For the cattle bounding box regression, only the features obtained from the original region proposal are used. Meanwhile the mask R-CNN head part predicts segmentation masks (cattle body regions).

For the network training, the loss function indicates the difference between the predicted value and the ground truth. It plays an important role in the cattle segmentation model training. In our Mask R-CNN based cattle segmentation network, a joint loss function is used to train bounding box regression, classification, and mask prediction branches. The used loss function is as follows:

$$L = L_{cls} + L_{bbox} + L_{mask} \tag{5}$$

where L_{cls} indicates the classification error; L_{bbox} is the bounding box regression error; L_{mask} indicates the mask error.

The classification error L_{cls} is computed by:

$$L_{cls} = \frac{1}{N_{cls}} \sum_i -\log[p_i^* p_i + (1 - p_i^*)(1 - p_i)] \tag{6}$$

where N_{cls} indicates the number of categories; p_i is the probability that the *i*-th RoIs are predicted to be positive samples (foreground). When the RoIs are positive samples, $p_i^* = 1$; otherwise, $p_i^* = 0$.

The equation for bounding box regression error L_{bbox} computation is as follows:

$$L_{bbox} = \frac{1}{N_{reg}} \sum_i p_i^* R(t_i, t_i^*)) \tag{7}$$

where N_{reg} is the pixel number in the feature map, t_i indicates four translation scaling parameters of positive sample RoIs to the prediction region; t_i^* indicates the four translation scaling parameters of positive sample RoIs to the real label; $R(\cdot)$ is a smooth function.

The mask error L_{mask} can be obtained by:

$$L_{mask} = -\frac{1}{m^2} \sum_{1 \leqslant i,j \leqslant m} [y_{ij}^k = (1 - y_{ij})\log(1 - y_{ij}^2)] \tag{8}$$

where y_{ij} is label value of the coordinate point (i, j) in the $m \times m$ region and y_{ij}^k is the predicted value for the *k*th class at that point.

3.6. Cattle contour extraction

Cattle contour is helpful in obtaining body parameters (Gomes et al., 2016). In our work, cattle contour was extracted based on the segmented cattle images.

Each segmented image is transformed to a binary matrix, where forehead (cattle image) is scored as '1' and background pixels are regarded as '0'. The cattle contour line should be the '1' pixels which are surrounded by the background pixels. Using this method, the cattle contour line can be obtained.

4. Performance evaluation

To evaluate the performance of cattle segmentation and contour extraction, Mean Pixel Accuracy (MPA) and Average Distance Error (ADE) are used respectively. The MPA is a very popular measurement tool in the evaluation of image segmentation, and it is derived from the correctly segmented pixel:

$$\mathrm{MPA} = \frac{1}{k}\sum_{i=0}^{k}\frac{p_{ii}}{\sum_{j=0}^{k} p_{ij}} \tag{9}$$

where k is the total number of categories including the background, P_{ij} is the total number of the pixel whose real pixel class is i and predicted as j, and P_{ii} is the total number of pixel whose real pixel class is i but predicted as i.

For the extracted contour line evaluation, the ADE is computed as follows:

$$\mathrm{ADE} = \frac{A_{union} - A_{overlap}}{T_{contour}} \tag{10}$$

where A_{union} is the area encompassed by both the predicted mask and the ground-truth, $A_{overlap}$ is the overlapping area between the predicted mask and the ground-truth, and $T_{contour}$ is the pixel number of the ground-truth contour line.

5. Experiments

5.1. Experiment setup

In this study, TensorFlow (Abadi et al., 2016) with GPU is used to construct the cattle segmentation model. The details of hardware information in this experiment are illustrated in Table 1.

In our whole dataset, a total of 1188 key cattle images (with significant movement) were selected and their size was reduced to 960 × 540. In addition, according to whether conduct image enhancement or not, we have the corresponding raw and enhanced image datasets. In each dataset, the first 988 images are used as the training set and the other 200 images constitute the testing set independently. For the model training requirement, the LabelMe tool (Russell et al., 2008) was used to label the cattle area manually to produce our ground-truth.

In terms of the Mask R-CNN segmentation approach, ResNet-101 and Faster R-CNN with ResNet were used as backbone and head architecture respectively. In addition, pre-trained backbone weights based on the COCO dataset (Lin et al., 2014) were also adopted to accelerate the training process. Additionally, in order to verify the effectiveness of image enhancement, the proposed Mask R-CNN based segmentation models were trained on the raw and enhanced image datasets respectively. The total training process lasts almost 2 days with a 0.0003 learning rate.

5.2. Comparison with state-of-the-art methods

In addition, the proposed Mask R-CNN segmentation approach was also compared with two state-of-the-art instance methods—DeepMask (Pinheiro et al., 2015) and SharpMask (Pinheiro et al., 2016). DeepMask extracts image features and then passes these features to two fully connected layer branches—'mask' branch and the corresponding 'score' branch. After the two branches are jointly trained, the outputs from the score branch are used to select the best segmentation masks. The latter proposed SharpMask (Pinheiro et al., 2016) introduces a refinement module which iteratively keeps merging information from lower layers to the coarse segmentation produced. With this module, SharpMask can produce sharper, pixel-accurate object masks to refine object segmentation results based on low pixel-level information.

Table 1
The experimental hardware.

Hardware	Type
CPU	Intel Xeon E5-2630 @ 2.20 GHz × 20
Memory	32 GB
GPU	GeForce GTX 1080 Ti
Hard disk	1 TB

For the experiments of DeepMask and SharpMask, the open-source code[1] was used. Both these two networks were trained with a 0.001 learning rate. The whole training time is around 10 h for each network. Like the Mask R-CNN segmentation model, both DeepMask and SharpMask network were also trained on the raw and enhanced image datasets respectively.

5.3. Cattle image segmentation results

The segmentation results on the raw and enhanced image datasets can be seen in Fig. 6. It can be seen that the segmented cattle body area (red region) especially the regions nearby the cattle head and leg are obviously overstep the real cattle contour on the raw images in Fig. 6(a), while the overstepping situation in the leg and head regions is significantly reduced on the enhanced images as shown in Fig. 6(b). The main reason is that the used image enhanced method mitigating the influence of shadows and illumination variation, that the misjudgment pixels between the shadows and real cattle body are reduced.

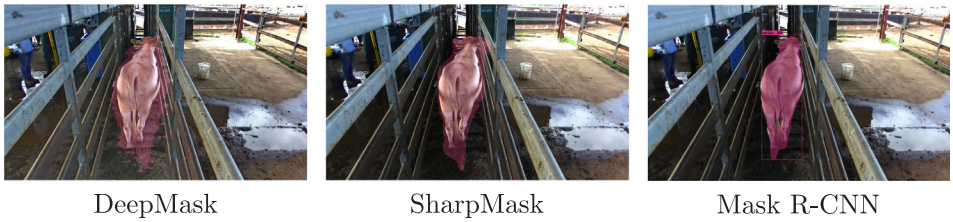
In addition, some multi-cattle instance segmentation results of the proposed Mask R-CNN approach are displayed in Fig. 7. It can be seen that even if there are two beef cattle in the crush, the proposed approach can still detect and segment each individual animal. It also can be noted that the proposed Mask R-CNN based cattle segmentation is not influenced by cattle coat colors (i.e. brown, white or dark) and poses.

The cattle segmentation accuracies and processing time are also given in Table 2. It can be found that the achieved MPA values on the enhance image dataset are approximate 2% higher than that on the raw image dataset. It illustrates that the improved image quality by the used enhancement method is helpful in the cattle segmentation. In addition, on the enhanced image dataset, the Mask R-CNN based cattle segmentation approach achieved 0.92 MPA with 0.73s processing time for each image, which is better than that of Deepmask (0.53 MPA) and SharpMask (0.82 MPA) on both the raw and enhanced image datasets. It illustrates again that the proposed Mask R-CNN based cattle segmentation on the enhanced image dataset is favorable.

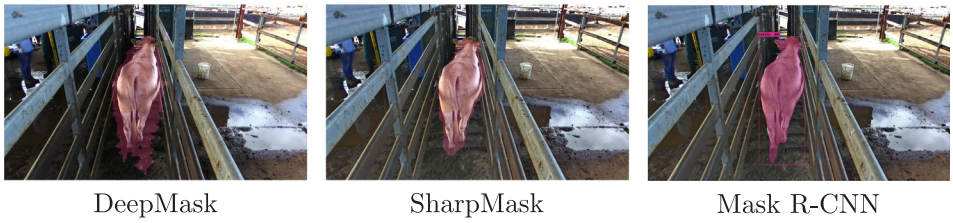
5.4. Cattle contour extraction results

For most cattle welfare measurements, the body parameters (e.g. body width and length) extracted from cattle contours are necessary and important (Van Hertem et al., 2013). Based on the segmentation results of the Mask R-CNN based approach, individual cattle contour line can be achieved easily. Since the segmentation results on the

[1] https://github.com/aby2s/sharpmask.



(a) Segmentation on raw images



(b) Segmentation on enhanced images

Fig. 6. Comparison of cattle image segmentation results on raw images (top row) and enhanced images (bottom row). From left to right column, the corresponding segmentation methods are DeepMask, SharpMask and Mask R-CNN respectively.

(a) Enhaced images

(b) Mask R-CNN segmentation

Fig. 7. The proposed Mask R-CNN based multi-cattle segmentation results. Examples of the enhanced images are displayed in the top row and their corresponding Mask R-CNN segmentation results are illustrated in the bottom row. Each color area indicates a segmented cattle instance. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 2
Accuracies of cattle image segmentation.

Methods	Data type	MPA	Time (s)
DeepMask	Raw	0.37	1.27
	Enchanced	0.53	1.27
SharpMask	Raw	0.79	1.34
	Enchanced	0.82	1.34
Mask R-CNN	Raw	0.90	0.73
	Enchanced	**0.92**	0.73

The best value is marked in bold.

enhanced image dataset are better than those of the raw image dataset, in this section, cattle contour line extraction experiments were conducted only on the enhanced image dataset.

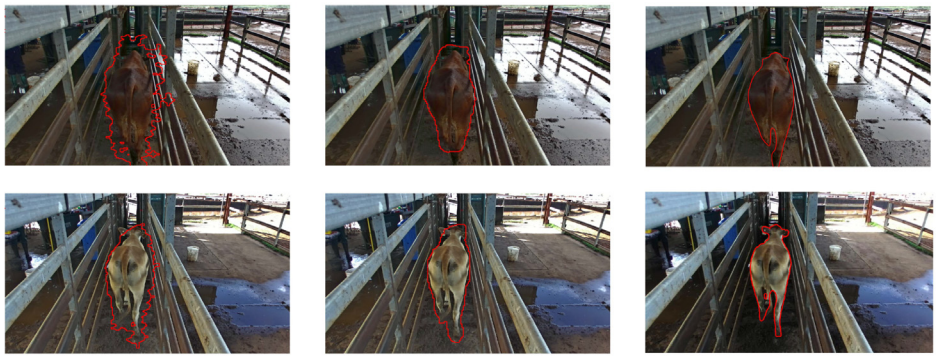
The contour line extraction results of DeepMask, SharpMask and Mask R-CNN are illustrated in Fig. 8. It can be seen that the performance of contour line extraction based on Mask R-CNN segmentation is very close to the real cattle contour, which is better than that of DeepMask and SharpMask (the extracted contour overstep the real cattle contour).

More cattle contour extraction examples of the Mask R-CNN based approach can be seen in Fig. 9. The top row in Fig. 9 shows some good contour extraction examples, while the bottom row displays some negative cases. It can be found that the overall performance of cattle contour extraction is favorable except for few negative cases in the extreme dark or the heavily overlapping situations. It is believed that, with more training data, the performance of segmentation and contour extraction can be further improved.

By comparing with the ground-truth (manually labeling using LabelMe tool), the Average Distance Error (ADE) of the extracted contour line is given in Table 3. It shows that Mask R-CNN based approach achieved 33.56 ADE of the extracted contour line, which is significantly better than those of the DeepMask (53.05 ADE) and SharpMask (42.64 ADE). It illustrates again that the proposed Mask R-CNN approach with the image enhancement is effective for cattle segmentation and contour extraction in the complex feedlot environment.

6. Conclusions

A robust and real-time cattle segmentation and contour extraction method is essential to further cattle welfare evaluation in precision livestock management. In order to realize cattle segmentation and contour extraction in complex feedlot environments (e.g. illumination variation and dynamic background), an effective Mask R-CNN based approach is proposed. The proposed approach selects key frames from



(a) DeepMask (b) SharpMask (c) Mask R-CNN

Fig. 8. Comparison of contour extraction results on enhanced images. The red line in each image is the extracted contour. Form left to right column, the corresponding methods are DeepMask, SharpMask and Mask R-CNN respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Fig. 9. Examples of cattle contour extraction using the proposed Mask R-CNN method on enhanced image datasets.

Table 3
Cattle contour extraction results.

Methods	Min	Max	ADE	Time(s)
DeepMask	0.052	112.54	53.05	1.29
SharpMask	0.074	68.71	42.64	1.35
Mask R-CNN	0.035	64.17	**33.56**	0.77

The best value is marked in bold.

image sequence to monitor the cattle movement, improves the image quality by reducing the shadows and illumination influence using a two-dimensional gamma function based image adaptive correction algorithm, and then realizes high-accuracy automatic segmentation and contour extraction. According to the experimental results on our acquired cattle dataset, Mask R-CNN based approach achieved MPA is 0.92 on the enhanced image dataset, which is 2 percent higher than that of the raw image dataset. Based on the Mask R-CNN segmentation, the achieved ADE of extracted cattle contour is 33.56 on the enhanced image dataset. In addition, for either the segmentation or contour extraction, the proposed Mask R-CNN based approach performs better than the state-of-the-art methods (DeepMask and SharpMask).

For future research, we intend to further improve the segmentation ability for the overlapping cattle regions. Additionally, explicit segmentation to distinguish different cattle body parts (e.g. head, trunk or leg) will also be considered.

Acknowledgments

The authors express their gratitude to Australian Country Choice and the Brisbane Valley farm staff for their help with data collection. Also particular thanks to Sabrina Lomax, Cameron Clark, Amanda Doughty, Ashraful Islam and Mike Reynolds for their involvement and efforts in the whole experiment organization and cattle information collection. The authors also acknowledge the support of the Meat & Livestock Australia Donor Company through the project: Objective, robust, real-time animal welfare measures for the Australian red meat industry. In addition, thanks to He Kong for his support in the manuscript revision.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.compag.2019.104958.

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al., 2016. Tensorflow: a system for large-scale machine learning. In: OSDI, vol. 16. pp. 265–283.

Andrew, W., Greatwood, C., Burghardt, T., 2017. Visual localisation and individual identification of holstein friesian cattle via deep learning. In: Proc. IEEE International Conference on Computer Vision (ICCV), Venice, Italy, pp. 22–29.

Bai, M., Urtasun, R., 2017. Deep watershed transform for instance segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5221–5229.

Bos, J.M., Bovenkerk, B., Feindt, P.H., Van Dam, Y.K., 2018. The quantified animal: precision livestock farming and the ethical implications of objectification. Food Ethics 2, 77–92.

Chen, L.-C., Hermans, A., Papandreou, G., Schroff, F., Wang, P., Adam, H., 2018. Masklab: Instance segmentation by refining object detection with semantic and direction features. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4013–4022.

Gardenier, J., Underwood, J., Clark, C., 2018. Object detection for cattle gait tracking. In: 2018 IEEE International Conference on Robotics and Automation (ICRA). IEEE, pp. 2206–2213.

Girshick, R., 2015. Fast R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1440–1448.

Girshick, R., Donahue, J., Darrell, T., Malik, J., 2015. Region-based convolutional networks for accurate object detection and segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 38, 142–158.

Gomes, R., Monteiro, G., Assis, G., Busato, K., Ladeira, M., Chizzotti, M., 2016. Estimating body weight and body composition of beef cattle trough digital image analysis. J.

Anim. Sci. 94, 5414–5422.
Halachmi, I., Guarino, M., 2016. Precision livestock farming: a per animal approach using advanced monitoring technologies. Animal 10, 1482–1483.
Hansen, M.F., Smith, M.L., Smith, L.N., Jabbar, K.A., Forbes, D., 2018. Automated monitoring of dairy cow body condition, mobility and weight using a single 3d video capture device. Comput. Ind. 98, 14–22.
Hariharan, B., Arbeláez, P., Girshick, R., Malik, J., 2014. Simultaneous detection and segmentation. In: European Conference on Computer Vision. Springer, pp. 297–312.
He, D., Liu, D., Zhao, K., 2016. Review of perceiving animal information and behavior in precision livestock farming. Trans. Chinese Soc. Agric. Mach. 47, 231–244.
He, K., Gkioxari, G., Dollár, P., Girshick, R., 2017. Mask R-CNN. In: 2017 IEEE International Conference on Computer Vision (ICCV). IEEE, pp. 2980–2988.
Kashiha, M., Bahr, C., Ott, S., Moons, C.P., Niewold, T.A., Ödberg, F.O., Berckmans, D., 2014. Automatic weight estimation of individual pigs using image analysis. Comput. Electron. Agric. 107, 38–44.
Krukowski, M., 2009. Automatic determination of body condition score of dairy cows from 3D images. Skolan för datavetenskap och kommunikation, Kungliga Tekniska högskolan.
Kumar, S., Pandey, A., Satwik, K.S.R., Kumar, S., Singh, S.K., Singh, A.K., Mohan, A., 2018. Deep learning framework for recognition of cattle using muzzle point image pattern. Measurement 116, 1–17.
Land, E.H., 1986. An alternative technique for the computation of the designator in the retinex theory of color vision. Proc. Nat. Acad. Sci. 83, 3078–3080.
Li, K., Hariharan, B., Malik, J., 2016. Iterative instance segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3659–3667.
Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S., 2017. Feature pyramid networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2117–2125.
Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L., 2014. Microsoft coco: common objects in context. In: European Conference on Computer Vision. Springer, pp. 740–755.
Liu, Z., Wang, D., Liu, Y., Liu, X., 2016. Adaptive adjustment algorithm for non-uniform illumination images based on 2D gamma function. Trans. Beijing Inst. Technol. 36, 191–196.
Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440.
Lynn, N.C., Zin, T.T., Kobayashi, I., 2017. Automatic assessing body condition score from digital images by active shape model and multiple regression technique. In: Journal of Robotics, Networking and Artificial Life, International Conference on Artifical Life and Robotics, pp. 311–314.
Nasirahmadi, A., Edwards, S.A., Sturm, B., 2017. Implementation of machine vision for detecting behaviour of cattle and pigs. Livestock Sci. 202, 25–38.
Nilsson, M., Herlin, A., Ardö, H., Guzhva, O., Åström, K., Bergsten, C., 2015. Development of automatic surveillance of animal behaviour and welfare using image analysis and machine learned segmentation technique. Animal 9, 1859–1865.
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. Nat. Acad. Sci. 115, E5716–E5725.
Oczak, M., Ismayilova, G., Costa, A., Viazzi, S., Sonoda, L.T., Fels, M., Bahr, C., Hartung, J., Guarino, M., Berckmans, D., et al., 2013. Analysis of aggressive behaviours of pigs by automatic video recordings. Comput. Electron. Agric. 99, 209–217.
Pinheiro, P.O., Collobert, R., Dollár, P., 2015. Learning to segment object candidates. In: Advances in Neural Information Processing Systems, pp. 1990–1998.
Pinheiro, P.O., Lin, T.-Y., Collobert, R., Dollár, P., 2016. Learning to refine object segments. In: European Conference on Computer Vision. Springer, pp. 75–91.
Qiao, Y., Cappelle, C., Ruichek, Y., Yang, T., 2019. Convnet and LSH-based visual localization using localized sequence matching. Sensors 19, 2439.
Ren, S., He, K., Girshick, R., Sun, J., 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems, pp. 91–99.
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.
Schoder, U., Staunfenbiel, R., 2006. Methods to determine body fat reserves in the dairy cow with special regard to ultrasonograpic measurement of backfat thickness. J. Dairy Sci. 89.
Sheena, C.V., Narayanan, N., 2015. Key-frame extraction by analysis of histograms of video frames using statistical methods. Procedia Comput. Sci. 70, 36–40.
Song, X., Bokkers, E., van der Tol, P., Koerkamp, P.G., van Mourik, S., 2018. Automated body weight prediction of dairy cows using 3-dimensional vision. J. Dairy Sci. 101, 4448–4459.
Stajnko, D., Brus, M., Hočevar, M., 2008. Estimation of bull live weight through thermographically measured body dimensions. Comput. Electron. Agric. 61, 233–240.
Tebug, S.F., Missohou, A., Sourokou Sabi, S., Juga, J., Poole, E.J., Tapio, M., Marshall, K., 2018. Using body measurements to estimate live weight of dairy cattle in low-input systems in senegal. J. Appl. Anim. Res. 46, 87–93.
Ter-Sarkisov, A., Ross, R., Kelleher, J., Earley, B., Keane, M., 2018. Beef cattle instance segmentation using fully convolutional neural network. arXiv preprint. 1807.01972.
Thornton, P.K., 2010. Livestock production: recent trends, future prospects. Philos. Trans. Roy. Soc. B: Biol. Sci. 365, 2853–2867.
Tscharke, M., Banhazi, T.M., 2016. A brief review of the application of machine vision in livestock behaviour analysis. Agrárinformatika/J. Agric. Informatics 7, 23–42.
Van Hertem, T., Alchanatis, V., Antler, A., Maltz, E., Halachmi, I., Schlageter-Tello, A., Lokhorst, C., Viazzi, S., Romanini, C., Pluk, A., et al., 2013. Comparison of segmentation algorithms for cow contour extraction from natural barn background in side view images. Comput. Electron. Agric. 91, 65–74.
Van Hertem, T., Rooijakkers, L., Berckmans, D., Fernández, A.P., Norton, T., Vranken, E., 2017. Appropriate data visualisation is key to precision livestock farming acceptance. Comput. Electron. Agric. 138, 1–10.
Wang, D., Tang, J., Zhu, W., Li, H., Xin, J., He, D., 2018. Dairy goat detection based on faster r-cnn from surveillance video. Comput. Electron. Agric. 154, 443–449.
Zhan, C., Duan, X., Xu, S., Song, Z., Luo, M., 2007. An improved moving object detection algorithm based on frame difference and edge detection. In: Image and Graphics, 2007. ICIG 2007. Fourth International Conference on. IEEE, pp. 519–523.
Zhang, A.L., Wu, B.P., Wuyun, C.T., Jiang, D.X., Xuan, E.C., Ma, F.Y., 2018. Algorithm of sheep body dimension measurement and its applications based on image analysis. Comput. Electron. Agric. 153, 33–45.
Zhao, K., Kang, J., Jung, J., Sohn, G., 2018. Building extraction from satellite images using mask r-cnn with building boundary regularization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 247–251.
Zin, T.T., Kobayashi, I., Tin, P., Hama, H., 2016. A general video surveillance framework for animal behavior analysis. In: 2016 Third International Conference on omputing Measurement Control and Sensor Network (CMCSN). IEEE, pp. 130–133.
Zitnick, C., Jojic, N., Kang, S.B., 2005. Consistent segmentation for optical flow estimation. In: Tenth IEEE International Conference on Computer Vision (ICCV'05), vol. 1. IEEE, pp. 1308–1315 volume 2.