A computer vision-based method for spatial-temporal action recognition of tail-biting behaviour in group-housed pigs

Dong Liu [a,b], *Maciej Oczak* [c,d,**], *Kristina Maschat* [c,e], *Johannes Baumgartner* [c], *Bernadette Pletzer* [c], *Dongjian He* [a,f,*], *Tomas Norton* [b,***]

[a] *College of Mechanical and Electronic Engineering, Northwest A&F University, Yangling, 712100, China*
[b] *M3-BIORES, KU Leuven, Leuven, Belgium*
[c] *Institute of Animal Welfare Science, The University of Veterinary Medicine Vienna (Vetmeduni Vienna), Veterinärplatz 1, 1210, Vienna, Austria*
[d] *Precision Livestock Farming Hub, The University of Veterinary Medicine Vienna (Vetmeduni Vienna), Veterinärplatz 1, 1210, Vienna, Austria*
[e] *FFoQSI GmbH, Technopark 1C, A-3430, Tulln, Austria*
[f] *Key Laboratory of Agricultural Internet of Things, Ministry of Agriculture and Rural Affairs, Yangling, Shaanxi, 712100, China*

As a typical harmful social behaviour, tail biting is considered to be a welfare-reducing problem with economic consequences for pig production. Taking a computer-vision based approach, in this study, we have developed a novel method to automatically identify and locate tail-biting interactions in group-housed pigs. The method employs a tracking-by-detection algorithm to simplify the group-level behaviour to pairwise interactions. Then, a convolution neural network (CNN) and a recurrent neural network (RNN) are combined to extract the spatial-temporal features and classify behaviour categories. The performance of the proposed method was evaluated by quantifying the localisation accuracy and behaviour classification accuracy. The results demonstrate that the tracking-by-detection approach is capable of obtaining the trajectories of biters and victims with a localisation accuracy of 92.71%. The spatial-temporal features trained by CNN and RNN are robust and effective with a category accuracy of 96.25%. In total, our proposed method is capable to identify and locate 89.23% of tail-biting behaviour in group-housed pigs.

Keywords:
Action recognition
Computer vision
Pig behaviour
Precision livestock farming
Tail biting

Nomenclature

A	Area ratio of two bounding boxes
AR	Aspect ratio of two bounding boxes
c	The confidence score of predicted bounding box
CS	Confidence score of the bounding box
D	Centre distance between two bounding boxes
g	Labelled bounding box (Ground truth)
$L(x, c, l, g)$	Loss function of the object detection model
$L_{conf}(x, c)$	Classification loss
$L_{loc}(x, l, g)$	Localisation loss
l	Predicted bounding box
MS	Matching score
N	The number of positive matches
x	An indicator for matching the predicted box to the ground truth box
α	Weight for localisation loss
α_1~α_4	Weights of tracking model indicators

Abbreviations

CNN	Convolution neural network
DNNs	Deep neural networks
Fast R-CNN	Fast Region-based convolution neural network
FN	False negative
FP	False positive
IoU	Intersection-over-union
LDA	Linear discriminant analysis
LSTM	Long short-term memory
*m*AP	Mean Average Precision
MHI	Motion history image
MTU	Minimum tracking unit
MS-COCO	Microsoft COCO dataset
OTA	Object tracking accuracy
OTP	Object tracking precision
PLF	Precision Livestock Farming
ResNet	Deep residual network
RNN	Recurrent neural network
SSD	Single-shot detector
TN	True negative
TP	True positive
VGG	Visual Geometry Group (A CNN architecture)
YOLO	You only look once (An object detection model)

1. Introduction

The pig industry and the general public have in recent years shown increased concern over pig health and welfare under intensive farming systems (Mellor, 2016). As a result, animal researchers have worked on establishing the relationship between changes in pig behaviour and their health and welfare status (Matthews, Miller, Clapp, Plötz, & Kyriazakis, 2016). Some research has focused on tail biting which is one of the most harmful behaviours and can result in reduction in pig welfare and in the profitability of pig production (Sonoda et al., 2013; Ursinus, Van Reenen, Kemp, & Bolhuis, 2014). The physical damage caused to the tail by a bite from another pig can lead to severe pain and secondary infection that can spread throughout the body of the victim pig. However, the reasons for tail biting are multifactorial, including external factors such as nutrition, group composition, and environment (floor type, stocking density, ventilation, and lack of enrichment) and internal factors such as genetics, sex and health status (Taylor, Main, Mendl, & Edwards, 2010; Zonderland et al., 2011). Therefore, it is difficult for farmers to design buildings that prevent tail biting on their farm. As a consequence, farmers must stay vigilant to reduce the deleterious effects of tail biting. However, monitoring of tail biting is still done manually and only when there is already blood visible in the pen and after an outbreak has already had significant consequences for the pigs. Such daily inspections of animals by farm stuff are laborious, subjective and come with high risks to animal health and productivity.

An early indication that an outbreak is likely to occur could help producers to intervene before major damage results. Therefore, tools that continuously and automatically monitor the risk of tail-biting outbreaks can be of high value to the industry. Previous studies have proved that indicators like tail-biting frequency, oral manipulation, tail position, and activity could be used predictors for an outbreak (Larsen, Andersen, & Pedersen, 2019; Statham, Green, Bichard, & Mendl, 2009; Taylor et al., 2010; Zonderland et al., 2011). Monitoring these variables with sensor technologies such as camera systems gives the possibility to realise early warning systems that can help manage tail biting (Banhazi et al., 2012). So far, very few studies have attempted to automatically identify tail-biting behaviour. In recent research, D'Eath et al. (2018) identified high/low tail position as a potential indicator of tail biting that was detectable by 3D cameras. They developed an algorithm to calculate the proportion of low to high tail positions within a group of pigs. However, tails at low positions were found to indicate more than just tail biting and position varied between groups and over time, e.g. the proportion of low tails increased when pigs were moved to a new pen. Therefore, a reliable methodology to detect tailing biting automatically is still missing.

Some related studies have been aimed at monitoring pig aggressive behaviour. To the author's knowledge, most of the existing methods assumed that the activity intensity of fighting pigs is much higher than other interactions. Based on this assumption, most studies design image features to quantify the changes of pixels. For example, Viazzi et al. (2014) used motion history image and occupation index to evaluate the activity intensity; Oczak et al. (2014) used the activity index and occupation index to evaluate the activity intensity; to eliminate the pixel changes caused by non-fighting pigs, Lee, Jin, Park, and Chung (2016) excluded the lying individuals by a height threshold in the depth image; Chen and Wang, et al. (2019) and Chen and Zhu, et al. (2019) used the area of the image connected area to exclude the motion of separated single pigs since the aggressive behaviour occurred between two pigs at least. Further, Chen et al. (2017) and Chen et al. (2018) directly obtained motion features from aggressive pigs by analysing connected area and adhesion index, achieving an accuracy of 97.04%. But this method is limited by the group scale, as the authors claimed that with the increase in the number of pigs in the limited space, the close contact between

pigs brings difficulties in locating aggressive pigs. Besides, regardless of the assumption of activity intensity and the limitation of the group scale, Chen et al. (2020) used a deep-learning-based method to recognise the aggressive video episodes of group-housed pigs without any hand-crafted heuristics.

Supplementary video related to this article can be found at https://doi.org/10.1016/j.biosystemseng.2020.04.007.

The main difference between tail-biting detection and pig aggression detection is that pigs displaying tail-biting behaviour will have no significant differences in activity intensity compared to the other pigs in the pen. The behaviour is usually expressed as an oral manipulation followed by the victims' reaction (through vocalisation and/or movement), both of which can be quite subtle. Hence, in order to detect small actions like tail biting, any behaviour features should be carried out at the individual level. While this poses significant challenges, addressing this problem also has the potential to lead to the development of a universal method that not only identifies a certain behaviour but also indicates from which pig the behaviour originated. This has significant benefits for industry and academia alike.

Therefore, this work aims to be the first study to develop a method for solving two key questions:

1). How to extract individual pigs' motion pattern in a group of pigs? We aim to design a tracking-by-detection algorithm that simplifies group-level pig behaviour to pairwise interactions.
2). How to accurately detect the tail-biting behaviour? We aim to integrate convolution neural networks (CNNs) and recurrent neural networks (RNNs) to monitor spatial-temporal features that are capable of classifying the tail-biting behavioural patterns.

2. Materials and methods

2.1. *Experimental materials and setup*

The data set for this study was collected during a feeding experiment at the research farm of the University of Veterinary Medicine Vienna (VetFarm Medau, Pottenstein, Lower Austria, Austria). The experiment was approved by the ethics committee of the University of Veterinary Medicine Vienna and the application for an animal experiment according to national legislation (§26 of the Law for Animal Experiments, Tierversuchsgesetz 2012) was accepted (68.205/0221-WF/V/3b/2017). Pigs ((Landrace × Large White) × Piétrain crossbreds) were housed in a fattening compartment of the testing unit of the research farm. The fattening compartment had 6 pens, each of area 19 m^2 (3.50 m × 5.48 m). Experiments were performed in 2 of the 6 pens. There were 12 fattening pigs housed in each of 2 pens. Thus, in total 24 undocked pigs from two fattening pens were selected as the research subjects. Pigs were fed *ad libitum* with one automatic feeding station ("Compident MLP" by Schauer Agrotronic GmbH, Prambachkirchen, Austria) per 12 animals. Water was provided permanently via nipple drinkers, 2 in each pen. Fattening pens had fully slatted floors and provided enrichment material of types and amounts complying with the Austrian animal welfare law. Two sisal ropes (length: 90 cm) with three knots in each were fixed to one wall of each experimental pen. They were renewed before they got so short the pigs could not take them in their mouths anymore. Furthermore, each pen was provided with a plastic hedgehog (Spieligel, Best Farm, GFS-Top-Animal-Service GmbH, Ascheberg, Germany), a plastic ball (Spielball groβ, 30 cm, befüllbar, GFS-Top-Animal-Service GmbH, Ascheberg, Germany) and a wooden beam hanging on a chain. The fattening compartment had an automatic ventilation system. The animals were introduced to the pens in July 2018 at the age of 10 weeks (30 kg) and stayed until slaughter (120 kg). To assist the labellers in identifying individual pigs, each pig was assigned an individual pattern that was drawn on their back with marking spray and renewed twice a week. However, this aspect was not used in the algorithms developed in this paper. Each fattening pen was equipped with an IP camera (GV-BX 1300KV, Geovision Inc., Taipei, Taiwan) locked in protective housing (HEB32K1, Videotec, Schio, Italy) hanging 5 m above the pen, giving an overhead view. Additionally, infrared spotlights (IR-LED294S-90, Microlight, Bad Nauheim, Germany) were installed in order to allow night recording. The images were recorded with 1280 × 720 pixels resolution in MPEG-4 format at 30 fps. The cameras were connected to a PC on which Multicam Surveillance System (8.5.6.0, Geovision Inc., Taipei, Taiwan) was installed. The system allowed simultaneous recording of videos from 6 cameras. Recordings were stored on exchangeable 3 TB hard drives. Eight hours of original 2D video recordings from two pens were selected for algorithm development. In pen 1, pigs (5 males, 7 females of approx. 35 kg) were recorded on day 8 after introduction to the pen, and pigs in pen 2 (6 males, 6 females of approx. 35 kg) on day 44. Animals in both pens were showing frequent tail-biting behaviour on the days of observation. All algorithms developed in this paper were implemented and trained in a computer equipped with Intel(R) i7-7700HQ CPU @2.80GHz, 16GB of RAM, and an NIVIDIA GTX 1050Ti GPU.

2.2. *Dataset description*

Two datasets were prepared for training two models. The first one, ***Pig Detection Dataset***, was used to train the object detection model (Section 2.3.1). The LabelImg software (https://github.com/tzutalin/labelImg) was used to label the location of each pig by giving a bounding box to each pig, as shown in Fig. 1. The software recorded four corner coordinates of each box. A total of 320 images with around 3000 pig instances were labelled and divided into training and validation sets, namely 256 images (80%) were randomly selected to train the object detection model and then the remaining 64 images (20%) were used to evaluate the performance.

The second dataset, ***Pig Action Dataset***, was used to train the action recognition model (Section 2.3.4). This dataset consisted of video segments of one-second length. Each video segment comprised an interaction between two pigs, including tail-biting and non-tail biting interactions.

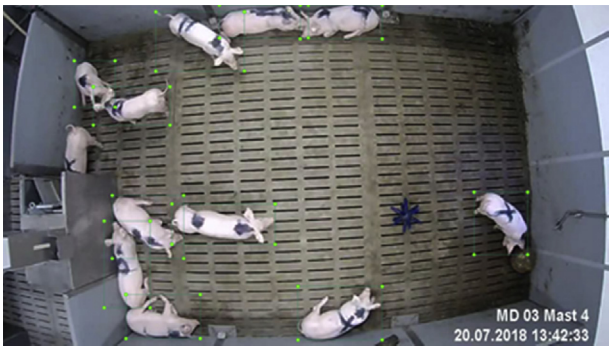

Fig. 1 – An example of labelled image in the Pig Detection Dataset. Every visible pig is manually labelled by giving a bounding box using LabelImg software.

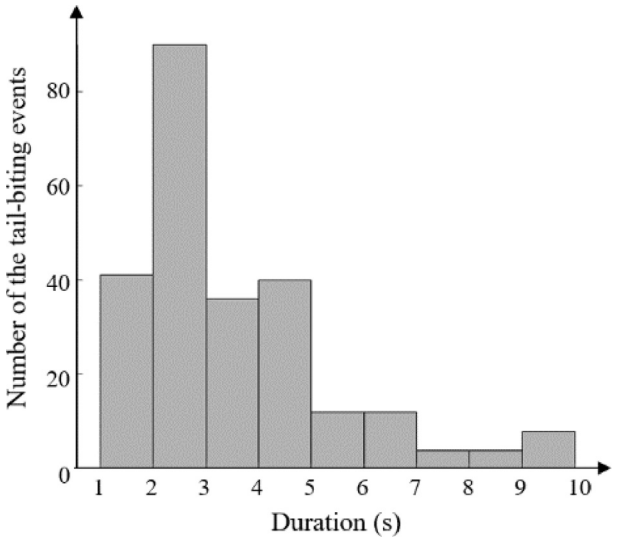

Fig. 2 – Frequency histogram of the duration of tail-biting events in this study.

To generate this dataset, first, the Interact software (version 9 and 14, Mangold International GmbH, Arnstorf, Germany) was used by a trained observer to label every tail-biting event from the original video. As described in Table 1, labelling of tail-biting events was done according to the definition of tail biting in the ethogram of Zonderland et al. (2011). When any two pigs in the group showed tail-biting behaviour, the start frame, end frame and duration of this biting event were recorded. The starting frame of a tail-biting event was determined by the time point at which the biter started to bite or chew the victim's tail, manifested as masticatory movements. The end frame was determined by the time point at which the victim's tail was released by the biter for more than one second. From the original video, a total of 247 biting events were recorded, lasting from 1s to 10s. Figure 2 shows the frequency histogram of the duration of tail-biting events.

Then, our proposed method (introduced in Section 2.3.1, Section 2.3.2 and Section 2.3.3) could automatically translate the original video into pairwise interactive sub-videos with a length of 1s. For example, a 10s tail-biting video could be divided into 10 sub-videos, being treated as 10 tail-biting events. The Pig Action Dataset was generated by adding the behaviour category information to these sub-videos. A total of 4396 interactions (742 sub-videos of tail-biting interactions and 3652 sub-videos of non-tail biting interactions) were extracted from the 247 tail-biting events. Note that the non-tail biting interactions cannot be equated to the pig ethogram (e.g. fighting, bullying, mounting, nosing, playing, and ear biting) in this study. These non-tail biting interactions were generated automatically by the proposed algorithm (Section 2.3), in which any two pigs that are close enough were recursively extracted as candidate interactions. As a result, around 90% of these "interactions" are just two pigs close together without any real interaction. The remaining meaningful interactions include chasing/following (198 cases), nosing (102 cases), pressing (58 cases), and head knocking (21 cases). No other violent interactions occurred in this dataset (e.g. fighting, bullying, or mounting). Since these meaningful interactions are too few to form a separate behaviour category, they were all considered as non-tail biting interactions in this dataset.

For the balance of the data set, data augmentation was applied to tail-biting interactions by rotating the image 90°, 180° and 270° respectively. Thus, 80% sub-videos (2374 tail-biting and 2921 non-tail biting sub-videos) were randomly selected as the training set and the remaining 20% sub-videos (594 tail-biting and 731 non-tail biting sub-videos) were used to test the performance of the action recognition model.

2.3. *Algorithms*

The goal of the proposed method was to detect and locate tail-biting interactions from group-housed pigs in the pen environment using computer vision technology. The basic idea to achieve this goal was to first simplify the group-level pigs' behaviour into the pairwise level. Then, an action recognition model was built based on the extracted pairwise interactions. The overall workflow of the framework is illustrated in Fig. 3. The first stage of the process aimed to extract sub-videos of

Table 1 – Tail-biting ethogram (Zonderland et al., 2011).

Tail biting	Description
Performed tail-biting behaviour (biter)	Biting a penmate's tail, with a sudden reaction of the penmate.
Received tail-biting behaviour (victim)	A penmate is biting the subject's tail and elicits a reaction.

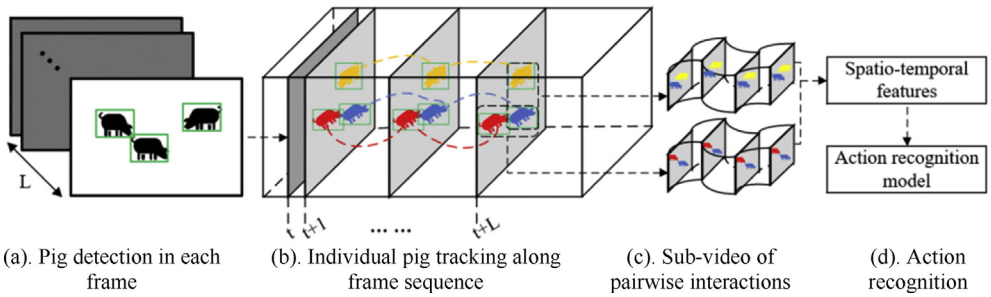

Fig. 3 – The overall workflow of the proposed framework. (a) Visible individual pigs were detected from the background. The location of each pig was represented by a bounding box. (b) Tracking was performed on each pig over L frames. (c) Sub-videos of pairwise interactions were extracted using the trajectory of each pig. (d) Action recognition model was built based on spatial-temporal features of these sub-videos.

pairwise interactions from the original video sequences. To do this, it was necessary to detect and track individual pigs along the video sequence. The following two sections (Section 2.3.1 and Section 2.3.2) describe the method used to detect and track each pig separately. Section 2.3.3 then explains steps to detect pairwise interactions based on the tracking results and the image pre-processing required to interface with the action recognition model that is later applied. Finally, Section 2.3.4 introduces the architecture of the action recognition model that takes pairwise-interaction video sequences as input, and then outputs the corresponding action category.

2.3.1. *Object detection*

Object detection is the process of finding instances of pigs in each frame. Thanks to the emergence of Deep Neural Networks (DNNs), established methods like YOLO (Redmon, Divvala, Girshick, & Farhadi, 2016), SSD (Liu et al., 2016), and Fast R-CNN (Girshick, 2015) could easily outperform all traditional methods for object detection (Zhao, Zheng, Xu, & Wu, 2019). Besides, one of the unexpected benefits of the object detection challenge (e.g. ImageNet (Russakovsky et al., 2015), MS-COCO (Lin et al., 2014)) is that deep architectures, trained on a big data set, can be used for other tasks and in other domains without it being necessary to train a model from scratch, so-called transfer learning (Pan & Yang, 2009). The pre-trained models are available from open source object detection toolboxes like Tensorflow Object Detection API (Huang et al., 2017) and MMDetection (Chen and Zhu, et al., 2019). These technologies provide an easy-to-implement solution for the purpose of locating individual animals from the group-housed environment (Ardö, Guzhva, Nilsson, & Herlin, 2018; Psota, Mittek, Pérez, Schmidt, & Mote, 2019). Actually, fine-tuning a pre-trained object detection model on a custom dataset has been successfully applied in Precision Livestock Farming (PLF). For example, Faster R-CNN was adapted to locate individual pigs in applications for recognition of lactating sow postures (Zheng et al., 2018) and feeding behaviour (Yang, Xiao, & Lin, 2018).

In this study, we assumed that images of pigs were captured from a top-down view, which means that the size and appearance of animals is consistent. Thus, SSD (Liu et al., 2016) is selected to locate visible pigs considering its real-time performance on detecting consistent targets. SSD is a one-stage object detection algorithm. This means that, in contrast to two-stage (e.g. Fast R-CNN (Girshick, 2015)), SSD combined two tasks (region proposal and prediction) into one network by utilising pre-defined boxes to look for objects. To predict the location of an object, one can establish a regression model between the pre-defined box and the labelled box (Fig. 1). To predict different sizes of the objects, the outputs of multiple convolutional layers were used to pre-define region candidates, as the deeper convolutional layers could cover larger receptive fields and more abstract representation. The original architecture of SSD used VGG-16 (Simonyan & Zisserman, 2014) as the base net (one of the widely used CNNs). In addition, Resnet-50 (He, Zhang, Ren, & Sun, 2016) is also introduced and replaced the base net (VGG-16) in this study. Since deeper networks are usually better for image classification, we assume that the feature map from Resnet-50 (50 layers) may be more distinguishable than that from VGG-16 (16 layers). Both CNNs (VGG-16 and Resnet-50) had been pre-trained on the MS-COCO dataset (Lin et al., 2014). The output of the object detection module includes location of each animal (coordination) and confidence score (0%~100%).

2.3.2. *Object tracking*

For the purpose of obtaining the motion pattern of the individual pig, it is necessary to determine the position of each pig along image sequences. In this study, a simple tracking-by-detection algorithm was proposed for the problem of multiple pig tracking, where the basic idea was to associate detections (bounding boxes) across frames. Ideally, if each pig in each frame can be properly located, the problem of multiple pig tracking is simplified to match these detections in each adjacent two frames. However, errors that lead to lost detections are always unavoidable due to the complexity of individuals' behaving in a group, which increased the risk of losing the object or tracking the wrong object. Therefore, in order to track the maximum number of pigs, an indicator called a Minimum Tracking Unit (MTU) was applied to this study with the aim to track stably in a short time. The MTU represents one second segments that were extracted from the

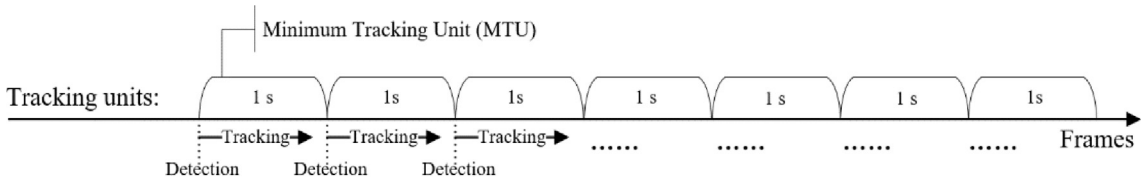

Fig. 4 – The schematic diagram of the minimum tracking unit.

original video. Figure 4 illustrates the schematic diagram of the minimum tracking unit. The reason for one second as an MTU was that we expected that the tracking unit can at least cover the shortest duration of one tail-biting event (according to the statistical analysis of labelled tail-bites in our dataset, as shown in Fig. 2). The tracking pipeline was independent of each MTU. The first frame of each MTU was used as the starting tracking point and the tracking process ended at the last frame of each MTU. Bounding-boxes with a confidence score of over 50% in the first frame were passed into the tracking pipeline.

The tracking pipeline was developed based on an assumption that the same object had a minor variance in the adjacent frames in terms of position and shape. The similarity of two boxes was measured by the Matching Score (MS), which is calculated by formula (1):

$$MS = CS \times \left(\alpha_1 \times {}^{1}/_{D} + \alpha_2 \times A + \alpha_3 \times AR + \alpha_4 \times IoU\right) \quad (1)$$

where, ***CS*** represents the confidence score of one candidate box in frame t+1 when searching for the best-matching box to a specific pig in frame t (0%~100%), ***D*** represents the centre distance (distance in number of pixels), ***A*** is the area ratio (0~1), ***AR*** represents aspect ratio (0~1), ***IoU*** represents intersection-over-union (0~1, Fig. 8) and α_1~α_4 are the weights of the corresponding indicator. Note that the Matching Score is normalised to a range between 0 and 1 in each MTU for consistency comparison, the larger the value, the higher the degree of approximation. Figure 5 illustrates the schematic diagram of the tracking pipeline in one MTU.

In this study, obtaining individuals' motion trajectory is crucial for the subsequent action recognition model, even it was acceptable to miss a few pigs with low matching confidence. To exclude those with obviously abnormal changes in adjacent frames usually caused by missing or misdetections (e.g. cases in Fig. 11), bounding boxes meeting the following rules were excluded from the tracking pipeline: (1) the category confidence is less than 50% (***CS*** < 0.5, e.g. the first image in Fig. 5); (2) the change in position is more than 25% (***IoU*** < 0.75, e.g. Figure 8); (3) the change in shape in more than 30% (***A*** < 0.7 or ***AR*** < 0.7); (4) the matching confidence in less than 50% (***MS*** < 0.5). Note that the position and the shape differences of a fast running pig in adjacent frames are even imperceptible by the naked eye.

If no box was matched to the target pig under the above conditions, then this one was considered as tracking failure and excluded from the current MTU. In addition, the object detector might lose one pig in a few frames and re-detect that pig in the following frames, resulting in lost tracking. To compensate for this kind of error, the tracking model searched for the best matching box frame by frame in the next 10 frames when there was no matching box in the adjacent frame. The third image in Fig. 5 shows an example, in which the object detector lost one pig (in red) while the tracking model found it in the fourth frame. Then, the lost pig was predicted by linear interpolation. The first image in Fig. 6 shows an example of the tracking result, in which seven out of ten pigs were tracked and visualised in different colours.

2.3.3. *Pairwise interactions*

As an interactive behaviour, tail biting must occur between two pigs. Section 2.3.1 and Section 2.3.2 provided approaches to obtain the trajectory of each pig within MTUs. To translate the individual's motion trajectory into the pairwise

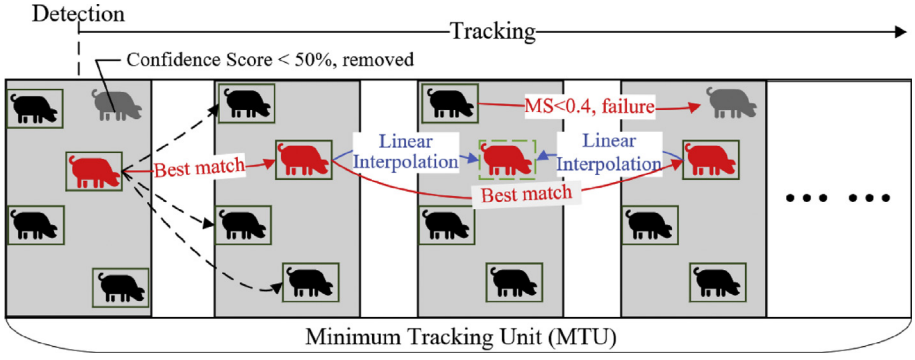

Fig. 5 – The schematic diagram of the proposed object tracking pipeline in one MTU.

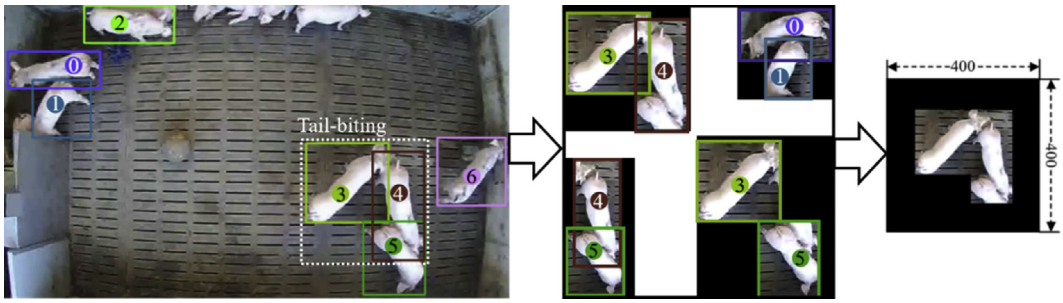

Fig. 6 – Extracting pairwise interactions from outputs of object tracking pipeline.

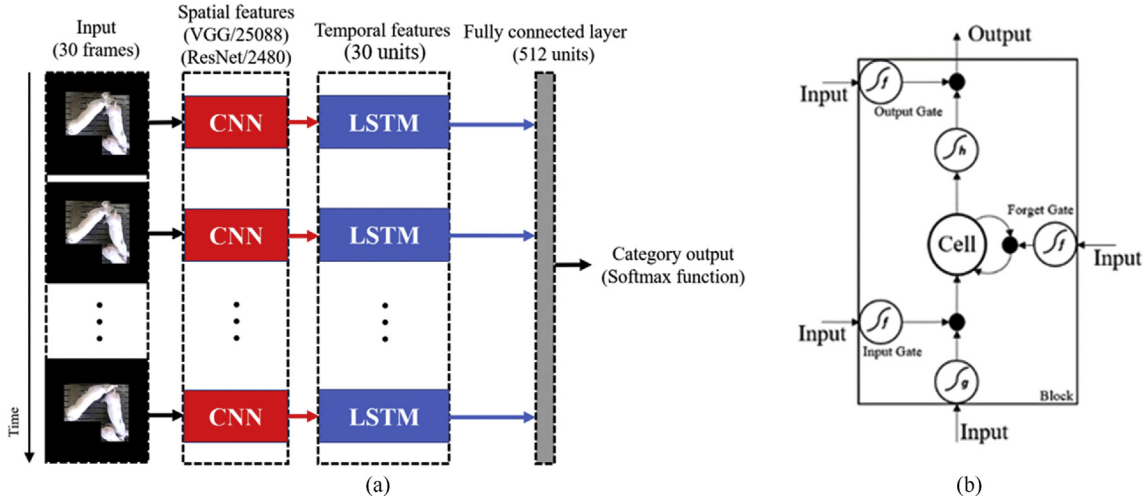

Fig. 7 – The architecture of the action recognition model. (a) The overall architecture. (b) A LSTM unit.

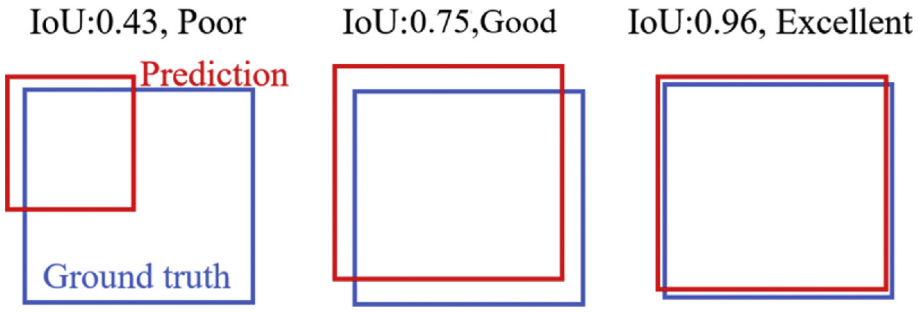

Fig. 8 – Example of how IoU is calculated. Predictions with IoU >0.75 were considered accurate in this study.

interactions, only two pigs' trajectories needs to be extracted while blocking all the others. This process needed to be repeated multiple times until it traversed all interactive pairs. Suppose that there are N pigs being tracked in one MTU, then N(N-1)/2 pairwise interactions can be extracted. Figure 6 shows an example where seven out of ten pigs are tracked successfully. A total of 21 pairwise interactions were obtained, where most of them are actually not "interactive behaviour" (e.g. Pig No.2 and Pig No.6). The distance of two bounding boxes can be used for filtering out those obvious non-interactive combinations. In the case shown in Fig. 6, only four valid interactions were left by applying the following two rules:

1). The average value of intersection-over-union (IoU) in an MTU should be greater than 0.

2). The margin distance of two bounding boxes should always be less than a threshold (30 pixels in this study). The margin distance is the external space separating two boxes.

Rule 1 operates on the frame sequences as a whole by calculating the average IoU but cannot ensure the two boxes are close enough at every moment, and Rule 2 is a supplement, making sure the boxes are close enough in an interactive process.

Besides, in order to obtain sub-videos with the same size, each extracted sub-frame is filled with 0 to a consistent shape (400*400 pixels).

2.3.4. *Action recognition*

A tail-biting interaction contains both spatial and temporal information. In the spatial domain (2D image), the biters' head must be close to the victims' tail (e.g. pig No. 3 with pig No.4 in Fig. 6). In the temporal domain (frame sequences), during a biting event, biters chase the escaping trajectory of victims (as described in Table 1). In this paper, both spatial and temporal features were learned without any hand-crafted heuristics. The architecture of the CNN + LSTM model is shown in Fig. 7a, including:

(1) The convolutional neural networks (CNNs) were employed to extract spatial features by encoding every 2D image into a 1D vector (convolutional feature). Then, the CNN features of every frame are combined into a sequence signal. Any position or posture changes in adjacent frames will result in differences in this sequence. Different sequence signals represent different types of interactions;

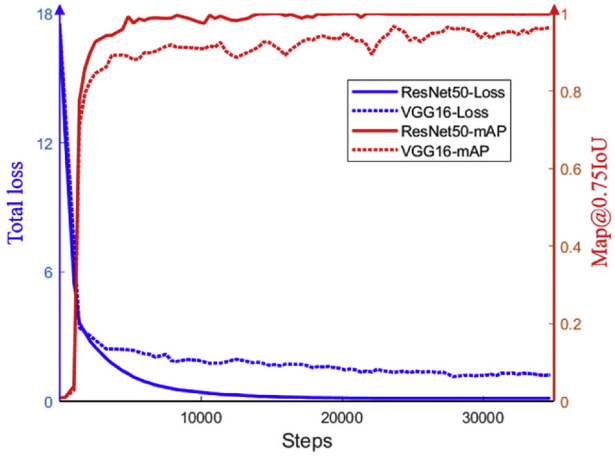

Fig. 9 – The metric for evaluating the performance of VGG-16 and ResNet-50. The total loss (blue in the left) and mAP curves (red in the right) are illustrated by 35,000 steps of training. In each step, number of Batch_Size samples are processed (Batch_Size = 12, Learning Rate = 0.04, gradient descent with momentum). (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

(2) The long short-term memory (LSTM) network (Hochreiter & Schmidhuber, 1997) receives this sequence signal as input. By training on the labelled Pig Action dataset, LSTM can distinguish the differences in the signal variation pattern. The architecture of LSTM unit is composed of one cell and three "regulators", namely input gate, output gate and forget gate (as shown in Fig. 7b). The cell is responsible for keeping track of the dependencies between the elements in the sequence signal. The input gate controls the extent to which a new value flows into the cell, the forget gate controls the extent to which a value remains in the cell and the output gate controls the extent to which the value in the cell is used to compute the output activation of the LSTM unit. The weights and activations of the "regulators" need to be learned during training, to make the cell remember values over arbitrary time intervals. The LSTM network was designed to classify, process and make predictions on entire sequences of data, such as speech recognition (Sak, Senior, & Beaufays, 2014) and action recognition (Donahue et al., 2015);

(3) A fully-connected neural net is concatenated at the end for categorical prediction.

As one of the hot topics in computer vision, action recognition has developed rapidly over the past few years. Deep-learning based approaches are highlighted (e.g. CNN + LSTM (Donahue et al., 2015), two-stream networks (Feichtenhofer, Pinz, & Zisserman, 2016), and 3D CNNs (Tran, Bourdev, Fergus, Torresani, & Paluri, 2015)). To the author's knowledge, there are so far few studies that attempt to test the feasibility of these models for animal behaviour. In this study, we did not focus on the state-of-the-art model. CNN + LSTM was chosen to verify the feasibility because its structure is relatively simple and easy to converge for a small-scale dataset. The CNN part can use pre-trained networks again, unlike other action recognition models that need to be trained from scratch. The LSTM part was trained from scratch. Besides, in order to verify whether the deeper CNN architecture had a better abstraction on spatial domain, ResNet-50 was also tested in addition to VGG-16.

3. Results

3.1. *The performance of object detection models*

The Pig Detection Dataset was used to train and validate the object detection model. The process of training the model is to minimise the loss function which is defined by formula (2):

$$L(x,c,l,g)=\frac{1}{N}\left(L_{conf}(x,c)+\alpha L_{loc}(x,l,g)\right) \quad (2)$$

where, $L_{loc}(x,l,g)$ measures the localisation loss between the predicted bounding box l and the ground truth box g; α is the weight for the localisation loss; $L_{conf}(x,c)$ is the loss in making a class prediction where the penalisation is conducted on the confidence score; and $x\in\{1,0\}$ is an indicator for matching the default box to the ground truth box (labelled box). If the

(a) (b)

Fig. 10 – Comparison of detecting results with different back-end CNN model. (a) Result of ResNet-based SSD. (b) Result of VGG-based SSD. Both models are able to detect all pigs, and ResNet-based SSD is better at confidence score.

IoU > 0.5 between the default box and the ground truth box, then $x = 1$, otherwise $x = 0$. N is the number of positive matches.

The performance can be evaluated by the mean Average Precision (*m*AP) (Everingham, Van Gool, Williams, Winn, & Zisserman, 2010), which is a comprehensive metric including category confidence scores (Average Precision, AP) and localisation precision (Intersection-over-Union, IoU). The higher the IoU, the more accurate the bounding box is (Fig. 8). The lower limit value of 0.75 was applied in this study because only high-accuracy locating made sense for subsequence tracking and behaviour recognition.

Figure 9 shows the loss curve and the mAP curve. The loss function drops rapidly and both models have been fitted at about 15,000 steps. This means that the transfer learning is feasible in this task, with some systematic similarities between the source domain (MS-COCO) and the target domain (Pig Detection Dataset). The object detection models learned how to extract the features during the pre-training phase. Then, the network further learned how to classify these features during the fine-tuning phase. As a result, transfer learning greatly improved the efficiency of training an object detection model as it normally takes weeks and millions of images to fit a deep model if training from scratch.

Figure 10 shows the output of the same sample in two models. Both of them can properly locate all pigs, and the ResNet-based model has a higher degree of confidence for some targets. The better performance of the ResNet-based SSD is mainly because ResNet provides skip-connections between convolutional blocks, thus diminishing the effects of vanishing gradient, allowing networks to go deeper. Normally, deeper CNNs help improve the performance of SSD, which is consistent with the conclusion from the image classification challenge (Russakovsky et al., 2015).

3.2. *The performance of tracking models and pairwise interception*

As a tracking-by-detection algorithm, the performance of the tracking model is heavily dependent on the detection results. In addition, once one pig can be properly tracked, its interactions with other pigs will certainly be extracted. The Object Tracking Precision (OTP) is utilised for the evaluation of tracker characteristic, which is computed by formula (3):

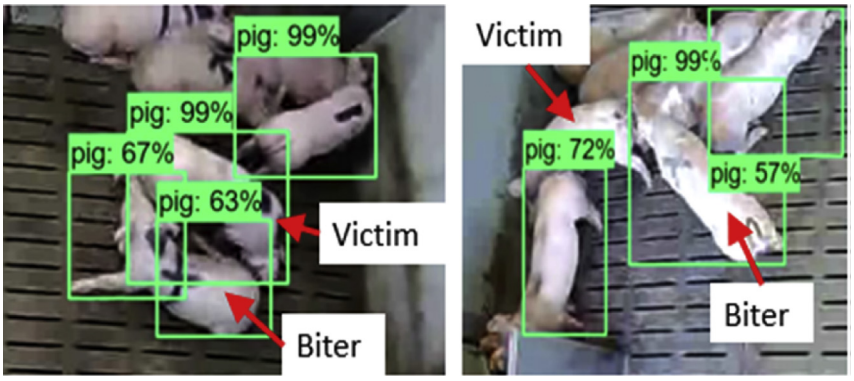

Fig. 11 – Fail cases of tracking tail-biting pigs. It is mainly because of errors in object detection.

$$OTP = \frac{\sum_{MTU_i} Boxes_{i,f=30}}{\sum_{MTU_i} Boxes_{i,f=1}} \quad (3)$$

This indicates the ability of the tracker to follow objects detected in the first frame within each MTU. $Boxes_{i,f=1}$ is the number of bounding boxes in the first frame of the ith MTU. $Boxes_{i,f=30}$ is the number of bounding boxes in the last frame of the ith MTU. Object Tracking Accuracy (OTA), indicates the ratio of properly tracked biting pigs accounted for all biting pigs (number of biters and victims being tracked/total number of biters and victims).

The performance of the tracking model is calculated in the case where the formula (1) has equal weights ($\alpha_1 \sim \alpha_4 = 0.25$). Table 2 summarises the results. Only 68.83% of detected pigs can be tracked properly, but 92.71% of tail-biting pigs are covered. The reasons behind the low OTP are: firstly, the camera had a fisheye effect, which caused the individuals at the image edge to deform. Unfortunately, pigs in this study preferred to crowd in the pen corner, probably to keep warm. Secondly, we deliberately did not label the individuals that were lying and occluded. But these individuals are occasionally detected due to the generalisation ability of the model. As a result, most of the failure cases are in group lying state, but fortunately, most of the tail-biting pigs could be successfully tracked. Figure 11 shows examples of fail cases. Note that OTP and OTA would drop to 52.35% and 71.23% respectively if only searching for the best match in the next one frame instead of ten frames (as described in Section 2.3.2). As the foundation of the framework, the most important component is to ensure that the tail-biting pigs can be detected and tracked precisely and this was achieved with an OTA of 92.71%.

Table 2 – Tracking model metrics.

Metrics	Data Set	Value
OTP (10 frames)	Pig Action Dataset (4396 samples)	68.83%
OTP (1 frames)		52.35%
OTA (10 frames)	247 tail-biting events (Fig. 1)	92.71%
OTA (1 frames)		71.23%

3.3. *Spatial-temporal features analysis*

In this section, the ResNet-based action recognition model was taken as an example to analyse the spatial-temporal features. The feature map (also called activation map/layers output) and the heat map were used to explain the construction of a behaviour feature vector by a video segment. The feature maps are the outputs of each CNN layer. They are obtained by performing convolution operation over the array of image pixels with a filter matrix. The heat map is the visualisation tool that interprets the patterns learned by convolutional layers (Selvaraju et al., 2017). For a particular feature map, the corresponding heat map could visualise which part of an image a CNN is looking at.

Figure 12 illustrates the process of extracting spatial features by ResNet-50. In the first layer, the feature map retained most of the information present in the original image and acted as edge detector, and the CNN focused on large regions with similar texture patterns. In deeper layers, the feature map showed a more abstract representation of the original image. From the output of the heat map, each feature map is an abstract representation of a specific image region. A total of 2048 feature maps from the 49th convolutional layer were used for spatial feature representation of the original image. They were then transferred into a 2048-dimension feature vector by average pooling in the last layer as the input to LSTM. Note that even if the pre-trained CNN has never seen our dataset before, it is still able to extract valid spatial features.

The LSTM learned temporal information by taking the convolution feature sequences as input. Figure 13 shows the changing pattern of the feature/heat maps in a time sequence. The feature map sequence showed significant sequence information, which is the result of the pigs' movement pattern. Meanwhile, the heat map pointed out that the feature maps actually track the displacement of a certain sub-region over time. For example, in Fig. 13a, the heat maps changed with the displacement of the two pigs which means the CNN always looked at the pigs' body. And the two pigs in Fig. 13b remained motionless, so the heat map had not changed.

Therefore, the CNN could extract the spatial features of the individual pig from image sequences, which made it possible

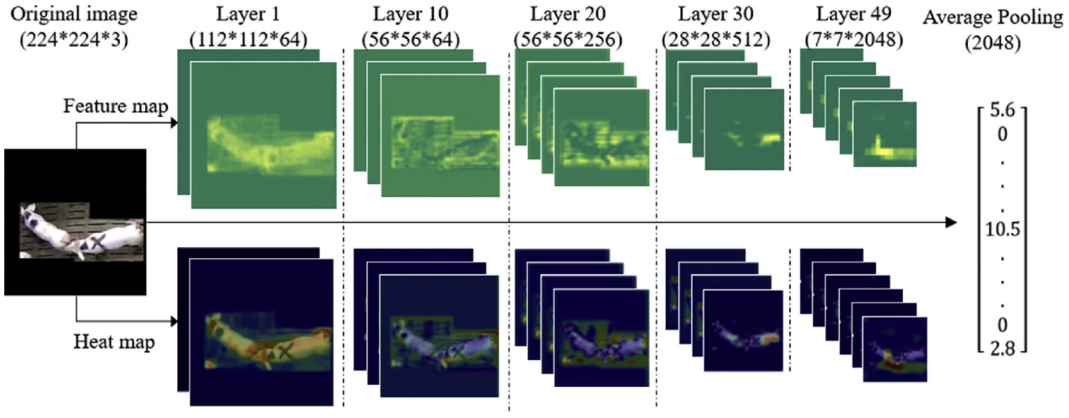

Fig. 12 – Feature map and heat map extracted by ResNet-50. The image size and the number of feature maps for each layer were presented by (Length*Width*Number).

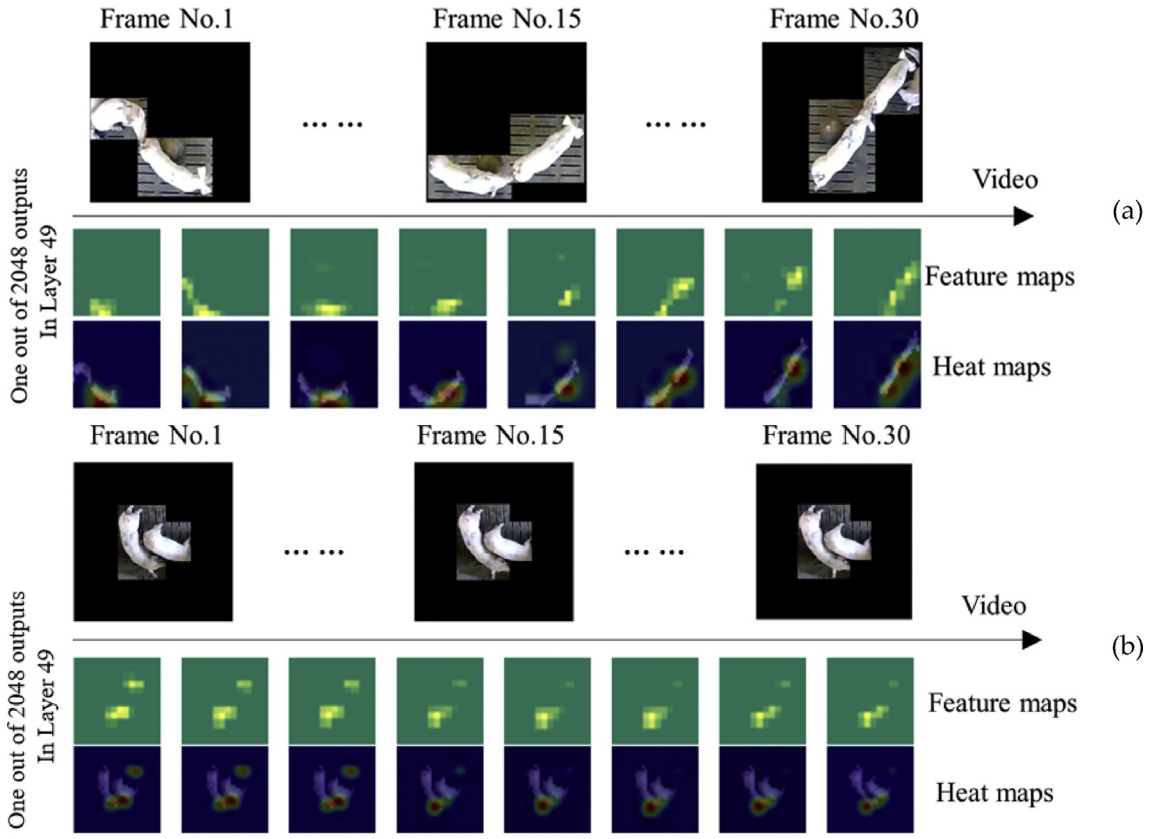

Fig. 13 – The changes of convolution features in the time sequence. (a) A tail-biting segment. (b) A non-tail biting segment.

to learn the motion pattern of tail-biting behaviour by training the LSTM network.

3.4. *Action recognition results*

The Pig Action Dataset was used to train and validate the action recognition models. Figure 14a, b show the accuracy and loss curves when training the CNN + LSTM model. Since the CNN model was frozen (act as spatial features extraction), only parameters of the LSTM were trained and that allowed the models to converge on the scale of data used in this study, otherwise a much larger data set might have been needed. With the same network architecture, the VGG + LSTM network converged to a validation accuracy of 92.24% after 20 iterations, but then over-fitted, and the ResNet + LSTM network converged to a validation accuracy of 96.35% after 60 iterations without over-fitting. The feature dimension of VGG-16 and ResNet-50 is 25,088 and 2480 respectively. Thus, for higher dimensional CNN features, it is necessary to either reduce the number of nodes of LSTM or enrich the training data.

In Fig. 14a, the validation accuracy can remain flat while the loss gets worse as long as the output value does not cross the threshold where the predicted class changes. A deep neural network is overfitted when progress in training is accompanied by increasing loss function and stable precision on the validation dataset. Vice versa, when the loss function decreases with a stable precision for the validation dataset, then the network capacity is weak and the model complexity needs to be increased.

Table 3 summarises the evaluation metrics of the ResNet-based action recognition result, including true positive (TP), true negative (TN), false positive (FP), false negative (FN), sensitivity ($=\frac{TP}{TP+FN}\times 100\%$), specificity ($=\frac{TN}{TN+FP}\times 100\%$), precision ($=\frac{TP}{TP+FP}\times 100\%$), and accuracy ($=\frac{TP+TN}{TP+TN+FP+FN}\times 100\%$). It can be seen that the proposed behaviour recognition model has good performance in recognition of the tail-biting interaction.

4. Discussion

The work in this paper provides a state-of-the-art method for monitoring tail-biting behaviour in a group. In comparison to previous research, in which tail posture was employed as an indirect indicator of tail-biting breakout (D'Eath et al., 2018), the current algorithm more effectively monitors the spatial-temporal characteristics of the tail-biting pattern by using deep-learning techniques. Comparing to related research, such as pig aggressive behaviour (Chen et al., 2017, 2018; Lee et al., 2016; Oczak et al., 2014; Viazzi et al., 2014), our approach not only extended monitoring to individual level but also enriched the capacity to capture better the

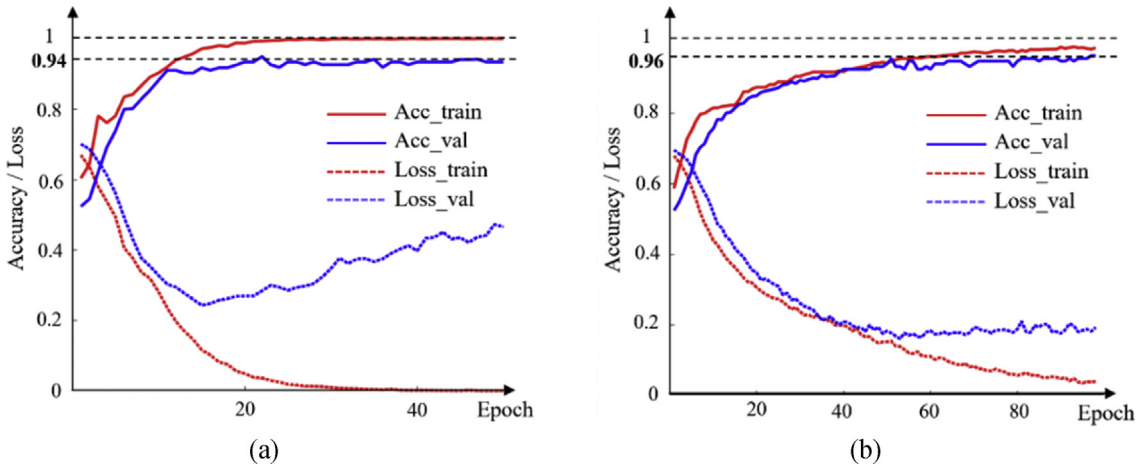

Fig. 14 – Process of training LSTM. (a) Training with VGG-16 convolution features. (b) Training with ResNet-50 convolution features. (Batch_Size = 12, Learning Rate = 0.01, gradient descent with Adam, Loss function with cross entropy).

Table 3 – The evaluation metrics of the ResNet-based action recognition model.

TP	TN	FP	FN	Sensitivity	Specificity	Precision	Accuracy
2288	2813	86	108	95.49%	97.03%	96.38%	96.35%

behavioural characteristics. Although, Chen et al. (2020) achieved an accuracy of 98.4% by submitting video episodes of the group-level aggressive behaviour into the CNN + LSTM model, the accuracy was only 52% when repeating the same method on tail-biting behaviour. This indicates that the CNN + LSTM model might not be sufficient to deal with the small actions that occur in the group. Therefore, the multiple object tracking algorithm is crucial to recognising tail-biting behaviour.

As a deep-learning based approach, 247 tail-biting events from 8 h raw video may seem insufficient to train a deep network. From an algorithm perspective, behaviour recognition on pairwise interaction level is a much more specific task than on the group level. By separating these pairs, the model only pays attention to the spatial-temporal features without considering the location. Thanks to the transfer learning techniques, a complex model can achieve good performance on the limited dataset. From the perspective of animal behaviour, providing toys in the pen will reduce the occurrence of tail-biting behaviour and thereby the size of the data sets. When collecting a large-scale animal behaviour dataset in the future, pens without toys should be considered.

There were three reasons for the failure in tail-biting location and recognition in this study:

(1) Loss of tail-biting pigs by the detecting or tracking model

There are several factors that may reduce the performance of the object detection model. One is the pigs' body change during a feeding cycle. Another one is the background changes of the pig pen (e.g. facilities and illumination). All of these problems could be addressed by enhancing the dataset. A large-scale video dataset is necessary for future research, covering the whole process of pig growth from day to night (with infrared camera), from nursery to fattening period, and crossing different farms and herd sizes. Increased diversity of the dataset can improve the generalisation ability of the object detector.

The tracking error actually originated from the object detection model. This could be improved by either eliminating the fisheye effect by correcting lens distortion, or utilising a depth sensor to completely remove the background. Moreover, the advanced object detection model (e.g. YOLO v3 (Redmon & Farhadi, 2018)) or an instance segmentation (e.g. DeepMask (Pinheiro, Collobert, & Dollar, 2015)) approach could be applied in the future.

(3) Misclassification of action recognition model

Figure 15 shows the typical examples of the classification result. It can be seen that:

(1) the relative position and the posture of interactive pigs are the key factor in determining behaviour category, since FP cases (False positive cases in Fig. 15a) show similar spatial characteristics to TP cases (Fig. 15c), and FN cases (Fig. 15b) have unusual postures compared to the typical tail-biting interaction (Fig. 15c);
(2) The main reason for FN recognition is when two pigs have a tail-biting interaction without obvious escaping and chasing motion (as shown in Fig. 15b);

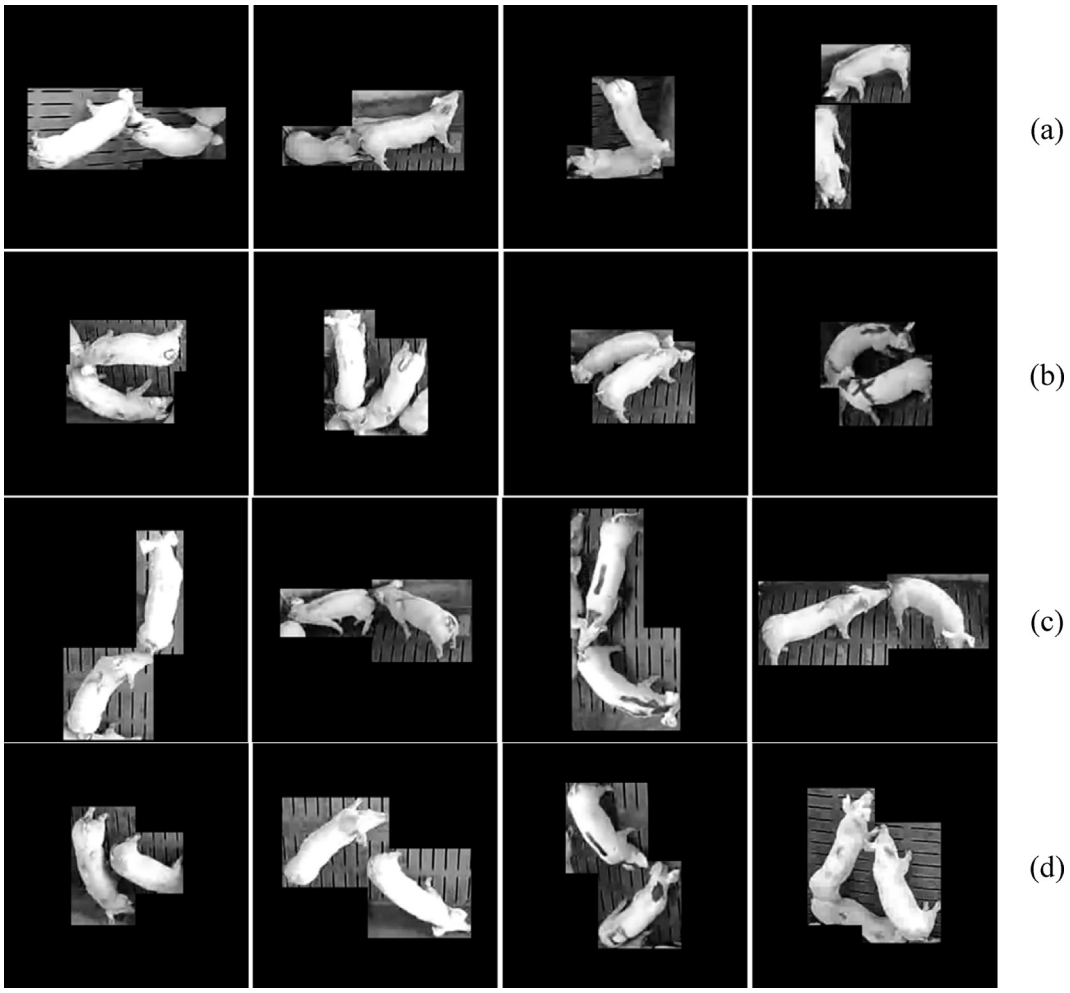

Fig. 15 – Typical examples of the behaviour recognition result. (a) Examples of the false positive; (b)Examples of the false negative; (c) Examples of the true positive; (c) Examples of the true negative.

(3) The main reason for FP recognition is when one pig wiggles its head around another pig's tail (as shown in Fig. 15a).

In conclusion, the CNN + SLTM model is applicable to tail-biting interactions with escaping and chasing motion. However, one of the tail-biting signals at a really early stage is that tail manipulation does happen while the pigs are lying down and does not necessarily invoke a reaction from the victim pig. The proposed method is defective in this respect according to the experimental results. A possible solution is that some novel action recognition models could learn spatial-temporal features synchronously, as this has been proved to be better than the CNN + LSTM model (e.g. two-stream networks (Feichtenhofer et al., 2016) and 3D CNNs (Tran et al., 2015)). However, due to the lack of a large-scale public dataset on pig behaviours, it is not possible to test these methods currently. Whether they are better than CNN + LSTM model or not needs to be verified in future work.

5. Conclusion

The method proposed in this paper attempted to provide a solution to locating and recognising typical harmful tail-biting behaviour in pigs. A major contribution of this method was to simplify group-level activities as pairwise interactions by a tracking-by-detection algorithm, which allowed precise localisation of the spatial position of the target interactions within the pig pen. Another contribution of this study was confirmation that the deep learning methods for human action recognition could also be feasible in animals. The proposed method has the potential to be used for monitoring multiple social behaviours in group-housed pigs.

Results demonstrate that the tracking-by-detection algorithm is capable of extracting 92.71% (229/247) tail-biting interactions from the raw video. Then, the accuracy of action recognition on the pairwise interactions reaches 96.25%. In conclusion, the proposed method can identify

and locate 89.23% of tail-biting interactions from group-housed pigs, which should meet the requirement of practical applications. This paper provides a solution to the early warning system of the tail-biting behaviour, which might help improve pig welfare and the profitability of pig industries.

Acknowledgment

This work was created within a research project of the Austrian Competence Centre for Feed and Food Quality, Safety and Innovation (FFoQSI). The COMET-K1 competence centre FFoQSI is funded by the Austrian ministries BMVIT, BMDW and the Austrian provinces Niederoesterreich, Upper Austria and Vienna within the scope of COMET -Competence Centers for Excellent Technologies. The programme COMET is handled by the Austrian Research Promotion Agency FFG. This research was also supported by the general program from the National Natural Science Foundation of China (Grant Number: 61473235), and the National Key Technology R&D Program of China (Grant Number: 2017YFD0701603). The authors would like to thank Barbara Metzler-Zebeli, Julia Vötterl, Jutamat Klinsoda and Thomas Enzinger from the Institute of Animal Nutrition and Functional Plant Compounds of University of Veterinary Medicine Vienna for excellent cooperation and practical support.

REFERENCES

Ardö, H., Guzhva, O., Nilsson, M., & Herlin, A. H. (2018). Convolutional neural network-based cow interaction watchdog. *IET Computer Vision, 12*, 171–177. https://doi.org/10.1049/iet-cvi.2017.0077.

Banhazi, T. M., Lehr, H., Black, J. L., Crabtree, H., Schofield, P., Tscharke, M., et al. (2012). Precision Livestock Farming: An international review of scientific and commercial aspects. *International Journal of Agricultural and Biological Engineering, 5*, 1–9. https://doi.org/10.3965/j.ijabe.20120503.001.

Chen, K., Wang, J., Pang, J., Cao, Y., Xiong, Y., Li, X., et al. (2019). *MMDetection: Open MMLab detection toolbox and benchmark*. arXiv preprint. arXiv:1906.07155.

Chen, C., Zhu, W., Guo, Y., Ma, C., Huang, W., & Ruan, C. (2018). A kinetic energy model based on machine vision for recognition of aggressive behaviours among group-housed pigs. *Livestock Science, 218*, 70–78. https://doi.org/10.1016/j.livsci.2018.10.013.

Chen, C., Zhu, W., Liu, D., Juan, S., Janice, S., Kaitlin, W., et al. (2019). Detection of aggressive behaviours in pigs using a RealSence depth sensor. *Computers and Electronics in Agriculture, 166*, 105003. https://doi.org/10.3390/s16050631.

Chen, C., Zhu, W., Ma, C., Guo, Y., Huang, W., & Ruan, C. (2017). Image motion feature extraction for recognition of aggressive behaviors among group-housed pigs. *Computers and Electronics in Agriculture, 142*, 380–387. https://doi.org/10.1016/j.compag.2017.09.013.

Chen, C., Zhu, W., Steibel, J., Siegford, J., Wurtz, K., Han, J., et al. (2020). Recognition of aggressive episodes of pigs based on convolutional neural network and long short-term memory. *Computers and Electronics in Agriculture, 169*, 105166.

Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., et al. (2015). Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR 2015), Boston, United States, 8–12 June 2015* (pp. 2625–2634). https://doi.org/10.1109/TPAMI.2016.2599174.

D'Eath, R. B., Jack, M., Futro, A., Talbot, D., Zhu, Q., Barclay, D., et al. (2018). Automatic early warning of tail biting in pigs: 3D cameras can detect lowered tail posture before an outbreak. *PloS One, 13*, e0194524. https://doi.org/10.1371/journal.pone.0194524.

Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., & Zisserman, A. (2010). The pascal visual object classes (VOC) challenge. *International Journal of Computer Vision, 88*, 303–338. https://doi.org/10.1007/s11263-009-0275-4.

Feichtenhofer, C., Pinz, A., & Zisserman, A. (2016). Convolutional two-stream network fusion for video action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR 2016), Las Vegas, United States, 26th June -1st July 2016* (pp. 1933–1941). https://doi.org/10.1109/CVPR.2016.213.

Girshick, R. (2015). Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision (ICCV 2015), Santiago, Chile, 7–13 December 2015* (pp. 1440–1448). https://doi.org/10.1109/ICCV.2015.169.

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 (CVPR 2016), Las Vegas, United States, 26th June -1st July 2016* (pp. 770–778). https://doi.org/10.1109/CVPR.2016.90.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*, 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735.

Huang, J., Rathod, V., Sun, C., Zhu, M., Korattikara, A., Fathi, A., et al. (2017). Speed/accuracy trade-offs for modern convolutional object detectors. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR 2017), Hawaii, United States, 21–26 July 2017* (pp. 7310–7311). https://doi.org/10.1109/CVPR.2017.351.

Larsen, M. L. V., Andersen, H. M. L., & Pedersen, L. J. (2019). Changes in activity and object manipulation before tail damage in finisher pigs as an early detector of tail biting. *Animal, 13*, 1037–1044. https://doi.org/10.1017/S1751731118002689.

Lee, J., Jin, L., Park, D., & Chung, Y. (2016). Automatic recognition of aggressive behavior in pigs using a kinect depth sensor. *Sensors, 16*, 631. https://doi.org/10.3390/s16050631.

Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., et al. (2014). Microsoft COCO: Common objects in context. In *Proceedings of the European conference on computer vision, Zurich, Switzerland, 6–12 September 2014* (pp. 740–755). https://doi.org/10.1007/978-3-319-10602-1_48.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., et al. (2016). SSD: Single shot multibox detector. In *Proceedings of the European conference on computer vision, Amsterdam, Netherlands, 8–16 October 2016* (pp. 21–37). https://doi.org/10.1007/978-3-319-46448-0_2.

Matthews, S. G., Miller, A. L., Clapp, J., Plötz, T., & Kyriazakis, I. (2016). Early detection of health and welfare compromises through automated detection of behavioural changes in pigs. *The Veterinary Journal, 217*, 43–51. https://doi.org/10.1016/j.tvjl.2016.09.005.

Mellor, D. J. (2016). Updating animal welfare thinking: Moving beyond the "five freedoms" towards "A lifeworth living". *Animals, 6*, 21. https://doi.org/10.3390/ani6030021.

Oczak, M., Viazzi, S., Ismayilova, G., Sonoda, L. T., Roulston, N., Fels, M., et al. (2014). Classification of aggressive behaviour in pigs by activity index and multilayer feed forward neural network. *Biosystems Engineering, 119*, 89–97. https://doi.org/10.1016/j.biosystemseng.2014.01.005.

Pan, S. J., & Yang, Q. (2009). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering, 22*, 1345–1359. https://doi.org/10.1109/TKDE.2009.191.

Pinheiro, P. O., Collobert, R., & Dollar, P. (2015). Learning to segment object candidates. In *Advances in neural information processing systems (NIPS 2015), Montreal, Canada, 7–12 December 2015* (pp. 1990–1998).

Psota, E. T., Mittek, M., Pérez, L. C., Schmidt, T., & Mote, B. (2019). Multi-pig part detection and association with a fully-convolutional network. *Sensors, 19*, 852. https://doi.org/10.3390/s19040852.

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 (CVPR 2016), Las Vegas, United States, 27–30 June 2016* (pp. 779–788). https://doi.org/10.1109/CVPR.2016.91.

Redmon, J., & Farhadi, A. (2018). *Yolov3: An incremental improvement.* arXiv Prepr. arXiv: 1804.02767.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., et al. (2015). ImageNet large scale visual recognition challenge. *International Journal of Computer Vision, 115*, 211–252. https://doi.org/10.1007/s11263-015-0816-y.

Sak, H., Senior, A., & Beaufays, F. (2014). Long short-term memory recurrent neural network architectures for large scale acoustic modeling. In *Fifteenth annual conference of the international speech communication association, Singapore, 14–18 September 2014* (pp. 338–342).

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision (ICCV 2017), Venice, Italy, 22–29 October 2017* (pp. 618–626). https://doi.org/10.1109/ICCV.2017.74.

Simonyan, K., & Zisserman, A. (2014). *Very deep convolutional networks for large-scale image recognition.* arXiv Prepr. arXiv:1409.1556.

Sonoda, L. T., Fels, M., Oczak, M., Vranken, E., Ismayilova, G., Guarino, M., et al. (2013). Tail biting in pigs—causes and management intervention strategies to reduce the behavioural disorder. A review. *Berl Munch Tierarztl Wochenschr, 126*, 104–112. https://doi.org/10.2376/0005-9366-126-104.

Statham, P., Green, L., Bichard, M., & Mendl, M. (2009). Predicting tail-biting from behaviour of pigs prior to outbreaks. *Applied Animal Behaviour Science, 121*, 157–164. https://doi.org/10.1016/j.applanim.2009.09.011.

Taylor, N. R., Main, D. C. J., Mendl, M., & Edwards, S. A. (2010). Tail-biting: A new perspective. *The Veterinary Journal, 186*, 137–147. https://doi.org/10.1016/j.tvjl.2009.08.028.

Tran, D., Bourdev, L., Fergus, R., Torresani, L., & Paluri, M. (2015). Learning spatiotemporal features with 3D convolutional networks. In *Proceedings of the IEEE international conference on computer vision (ICCV 2015), Santiago, Chile, 13–16 December 2015* (pp. 4489–4497). https://doi.org/10.1109/ICCV.2015.510.

Ursinus, W. W., Van Reenen, C. G., Kemp, B., & Bolhuis, J. E. (2014). Tail biting behaviour and tail damage in pigs and the relationship with general behaviour: Predicting the inevitable? *Applied Animal Behaviour Science, 156*, 22–36. https://doi.org/10.1016/j.applanim.2014.04.001.

Viazzi, S., Ismayilova, G., Oczak, M., Sonoda, L. T., Fels, M., Guarino, M., et al. (2014). Image feature extraction for classification of aggressive interactions among pigs. *Computers and Electronics in Agriculture, 104*, 57–62. https://doi.org/10.1016/j.compag.2014.03.010.

Yang, Q., Xiao, D., & Lin, S. (2018). Feeding behavior recognition for group-housed pigs with the Faster R-CNN. *Computers and Electronics in Agriculture, 147*, 51–63. https://doi.org/10.1016/j.compag.2018.11.002.

Zhao, Z. Q., Zheng, P., Xu, S. T., & Wu, X. (2019). Object detection with deep learning: A review. *IEEE Transactions on Neural Networks and Learning Systems, 30*, 3212–3232. https://doi.org/10.1109/TNNLS.2018.2876865.

Zheng, C., Zhu, X., Yang, X., Wang, L., Tu, S., & Xue, Y. (2018). Automatic recognition of lactating sow postures from depth images by deep learning detector. *Computers and Electronics in Agriculture, 147*, 51–63. https://doi.org/10.1016/j.compag.2018.01.023.

Zonderland, J. J., Schepers, F., Bracke, M. B. M., Den Hartog, L. A., Kemp, B., & Spoolder, H. A. M. (2011). Characteristics of biter and victim piglets apparent before a tail-biting outbreak. *Animal, 5*, 767–775. https://doi.org/10.1017/S1751731110002326.