Computers and Electronics in Agriculture 169 (2020) 105166

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

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

Recognition of aggressive episodes of pigs based on convolutional neural network and long short-term memory

Chen Chen[a,b], Weixing Zhu[a,*], Juan Steibel[c], Janice Siegford[c], Kaitlin Wurtz[c], Junjie Han[c], Tomas Norton[b,*]

[a] School of Electrical and Information Engineering, Jiangsu University, Zhenjiang 212013, Jiangsu, China
[b] Division of Measure, Model & Manage Bioresponses (M3-Biores), KU Leuven, Kasteelpark Arenberg 30, 3001 Leuven, Belgium
[c] Animal Behavior and Welfare Group, Department of Animal Science, Michigan State University, 3270C Anthony Hall, East Lansing, USA

ARTICLE INFO

Keywords:
Aggression recognition
Convolutional neural network
Long short-term memory
Deep learning
Computer vision

ABSTRACT

Aggression is considered as a major animal welfare problem in commercial pig farming. The aim of this study is to develop a deep learning method based on convolutional neural network (CNN) and long short-term memory (LSTM) to recognise aggressive episodes of pigs. Compared to previous studies of pig behaviours based on deep learning, this study directly process video episodes rather than individual frames. In the experiment, nursery pigs (8/pen) were mixed for 3 days and then 8 h of video was recorded in each day. From these videos, 600 aggressive 2 s-episodes were manually selected and then augmented into 2400 episodes by using horizontal, vertical and diagonal mirroring. From the videos, 2400 non-aggressive 2 s-episodes were also manually selected. 80% of the data were randomly allocated as training set and the remaining 20% as validation set. Firstly, the CNN architecture VGG-16 was used to extract spatial features. These features were then input into LSTM framework to further extract temporal features. Through fully connected layer, the prediction function *Softmax* was finally used to determine if the current episode is aggression or non-aggression. Using the proposed method, aggressive episodes could be recognised with an accuracy of 97.2%. This result indicates that this method can be used to recognise aggressive episodes of pigs. Additionally, this paper further investigates the validity of this method under the conditions of skipping frames and reducing the episode length. The results show that a frame skipping approach whereby 30 fps is reduced into 15 fps within each 2 s-episode can improve the accuracy into 98.4% and halve the total running time.

1. Introduction

In commercial pig production, the mixing of unacquainted pigs is a standard procedure that often results in aggressive interactions among pigs (Büttner et al., 2015). Aggression is considered as a major animal welfare problem in commercial pig farming (Peden et al., 2018), as it potentially leads to injury and stress and can influence the health and production of pigs (O'Malley et al., 2018). Compared to traditional manual observation of aggressive behaviours, computer vision technology has advantages, as it is non-intrusive, uninterrupted, can occur over longer periods and can be less-subjective. Therefore, there has been great interest in using computer vision for recognition of aggressive behaviours of pigs.

Initially, Viazzi et al. (2014) extracted the mean intensity and occupation index of the pig herd and then used linear discriminant analysis (LDA) to classify these 2 features to detect aggression. In order to further classify high and medium aggression, Oczak et al. (2014) extracted the average, maximum, minimum, sum and variance of activity index of the herd and then trained a multilayer feed forward neural network with these features to realise the classification. Subsequently, Lee et al. (2016) used a Kinect depth sensor to extract the maximum, minimum, average and standard deviation of the velocity of standing pigs and the distance between pigs as features. Support vector machine (SVM) was used to classify these features to detect aggression. Chen et al. (2017) further separated aggressive pigs from the herd by using connected area and adhesive index and then considered these two pigs as an entire rectangle to calculate its acceleration. Based on this acceleration, rules for aggression recognition were designed to recognise high and medium aggression. Chen et al. (2018) further located feature points of aggressive pigs on their contour and then extracted kinetic energy of these points. Kinetic energy differences between adjacent frames were used as features to recognise medium and high aggression.

* Corresponding authors.
E-mail addresses: wxzhu@ujs.edu.cn (W. Zhu), tomas.norton@kuleuven.be (T. Norton).

https://doi.org/10.1016/j.compag.2019.105166
Received 2 October 2019; Received in revised form 16 December 2019; Accepted 18 December 2019
Available online 26 December 2019

Additionally, Chen et al. (2019) further removed moving pixels caused by non-aggressive behaviours by setting a threshold of connected area in depth images and then summed the number of filtered moving pixels as motion shape index (MSI). The maximum, mean, variance and standard deviation of MSI in each 3 s-unit were extracted and classified by SVM in order to detect aggression.

However, the accuracy of the above studies is greatly affected by image quality and the amount of touching between pig bodies. These factors are influential because image quality affects the results of image segmentation and motion parameters, and the amount of touching between pig bodies affects ability to locate aggressive pigs. Relevant literature indicates that deep learning techniques can help to solve the problem of low image quality and touching pigs. For instance, Yang et al. (2018a) used a fully convolutional network (FCN) to segment images of lactating sows with different image qualities. Furthermore, Tian et al. (2019) used the modified Counting Convolutional Neural Network (CNN) model based on the architecture ResNeXt to count the number of pigs under the conditions of partial occlusion, overlapping and different perspectives. On the other hand, deep learning has also been used for recognition of other pig behaviours. For instance, Yang et al. (2018b) used a Faster R-CNN architecture to locate and identify individual pigs and then extracted feeding area occupation rate to recognise feeding behaviour. Zheng et al. (2018) used Faster R-CNN to recognise and classify lactating sow's 5 postures, i.e. standing, sitting, sterna recumbency, ventral recumbency and lateral recumbency. Additionally, Zhang et al. (2019) proposed a Sow Behaviour Detection Algorithm based on Deep Learning (SBDA-DL) to recognise drinking, urination and mounting of sows.

In the above studies of pig behaviours based on deep learning technology, deep learning was used to process images frame-by-frame. However, pig aggression is a fast and complex interactive behaviour and has the characteristics of continuous and large-proportion touching pig-bodies (McGlone, 1985; Chen et al., 2017). Therefore, this paper attempts to use a deep learning method with spatial-temporal information, i.e. recurrent neural network (RNN), to train and test video episodes in order to directly recognise whether aggression exists in these episodes or not. Long short-term memory (LSTM) is a type of commonly used RNN (Hochreiter and Schmidhuber, 1997), and it has been widely used for gesture recognition (Tsironi et al., 2017), online handwriting recognition (Nguyen, et al., 2018) and text report classification (Banerjee et al., 2019). The above applications indicate that LSTM has good performance in video recognition. Moreover, using CNN features extracted from images into LSTM networks can further obtain temporal information (Donahue et al., 2015; Srivastava et al., 2015). Hence, the objective of this study is to combine CNN and LSTM to extract spatial-temporal features in order to recognise aggressive episodes of pigs.

2. Materials and methods

2.1. Experimental setup

2.1.1. Video acquisition

The videos were collected from the swine teaching and research center of Michigan State University (East Lansing, MI, USA). Recently mixed nursery pigs were used in this study. The pigs were crossbred from a Yorkshire Dam x PIC composite male line. Each nursery pen was 1.78 m × 1.88 m, and it contained a self-feeder and a nipple drinker (Fig. 1). 1080p IR outdoor dome security camera (CTP-TLVA29AV, Cantek Plus, USA) was placed above the pen at the height of 2.44 m relative to the ground. This camera was used to record RGB videos with a resolution of 1180 × 830 pixels and a frame rate of 30 fps. At 26.5 ± 0.52 days of age, 2 pens each containing 8 castrated male pigs were created by mixing 4 pairs of familiar pigs together. Video was recorded for the first 3 days after mixing, and 8 h (09:00-17:00) was later decoded as experimental data. The rationale behind this approach to video acquisition is that the pigs express frequent aggressive behaviours during the first 3 days after mixing and thus sufficient data could be collected to meet the needs of the research (Erhard et al., 1997; Spoolder et al., 2000).

The computer processor was Intel(R) Core(TM) i7-8700 K CPU @ 3.70 GHz with 16 GB of RAM memory running a Microsoft Windows 10 Enterprise operating system. The graphic card was NVIDIA GeForce RTX 2080 with 8 GB of physical memory. The software used for developing algorithms was Python 3.7.3. CNN and LSTM were implemented on the frameworks of Tensorflow 1.13.1 and Keras 2.2.4, respectively.

2.1.2. Datasets and labelling

Through statistical analysis, it can be found that the minimum duration of aggressive behaviours was 2 s in this study. Therefore, 600 aggressive 2 s-episodes were manually selected from the recorded 3 days of videos. In order to augment aggressive data and maintain the image size, the transformation matrixes of horizontal, vertical and diagonal mirroring (i.e. [−1 0 0; 0 1 0; Width 0 1], [1 0 0; 0–1 0; 0 Height 1] and [−1 0 0; 0–1 0; Width Height 1]) and the *imwarp* function in MATLAB (R2018b, The MathWorks Inc., MA) were used to increase the number of aggressive episodes into 2400 (Fig. 1(a)). Among them, aggressive behaviours include head to head knocking, head to body knocking, parallel pressing, inverse parallel pressing, neck biting, body biting and ear biting. In order to keep the balance between the number of aggressive and non-aggressive episodes and increase the diversity of non-aggressive episodes, 10% of feeding, 10% of mounting, 10% of lying, 10% of chasing, 5% of running, 5% of escaping caused by being frightened, 10% of drinking and 40% of activities were manually selected to build 2400 non-aggressive 2 s-episodes (Fig. 1(b)). Among them, feeding, mounting and lying generate touching pigs. Chasing, running and escaping belong to high-speed motion, while drinking and activities belong to low-speed motion. The rationality of data allocation in this way is that the complexity of positive and negative samplings can be increased.

It is noted that LSTM randomly allocate training and validation sets. Namely, 80% of aggressive and non-aggressive 2 s-episodes were randomly selected as training set and the remaining 20% of 2 s-episodes were used as validation set. The specific steps are as followed:

1. Firstly, 80% of aggressive and non-aggressive data (i.e. 3840 2 s-episodes) and the remaining 20% of aggressive and non-aggressive data (i.e. 960 2 s-episodes) were respectively used as training and validation sets.
2. Then, the batchsize that represents the number of 2 s-episodes inputting in each time was set to 3. As a result, training set was divided into 1280 (=3840/3) units for training and iterations in order to obtain the minimum loss. The model generated by these 1280 iterations was used to test the 320 (=960/3) units in validation set to obtain the accuracy. This process is called as an epoch. After several epochs in turn, the loss of the model becomes smaller and smaller, and the accuracy gets higher and higher.
3. Finally, both the loss and the accuracy reach the optimal values (Fig. 6).

In order to further study the validity of the proposed method under less frames and smaller episode length to greatly reduce the training and test time, the *VideoReader*, *hasFrame* and *VideoWriter* functions in MATLAB were used to generate other 4 datasets based on the original dataset of 2 s_60 frames (Dataset 1). Dataset 2: The dataset of 2 s_30 frames was generated by skipping 1 frame in the video episodes of 2 s_60 frames. Dataset 3: The dataset of 2 s_20 frames was generated by skipping 2 frames in the video episodes of 2 s_60 frames. Dataset 4: The dataset of 1 s_30 frames was generated by using the video episodes of the first half of 2 s_60 frames. Dataset 5: The dataset of 1 s_15 frames was generated by skipping 1 frame in the video episodes of 1 s_30

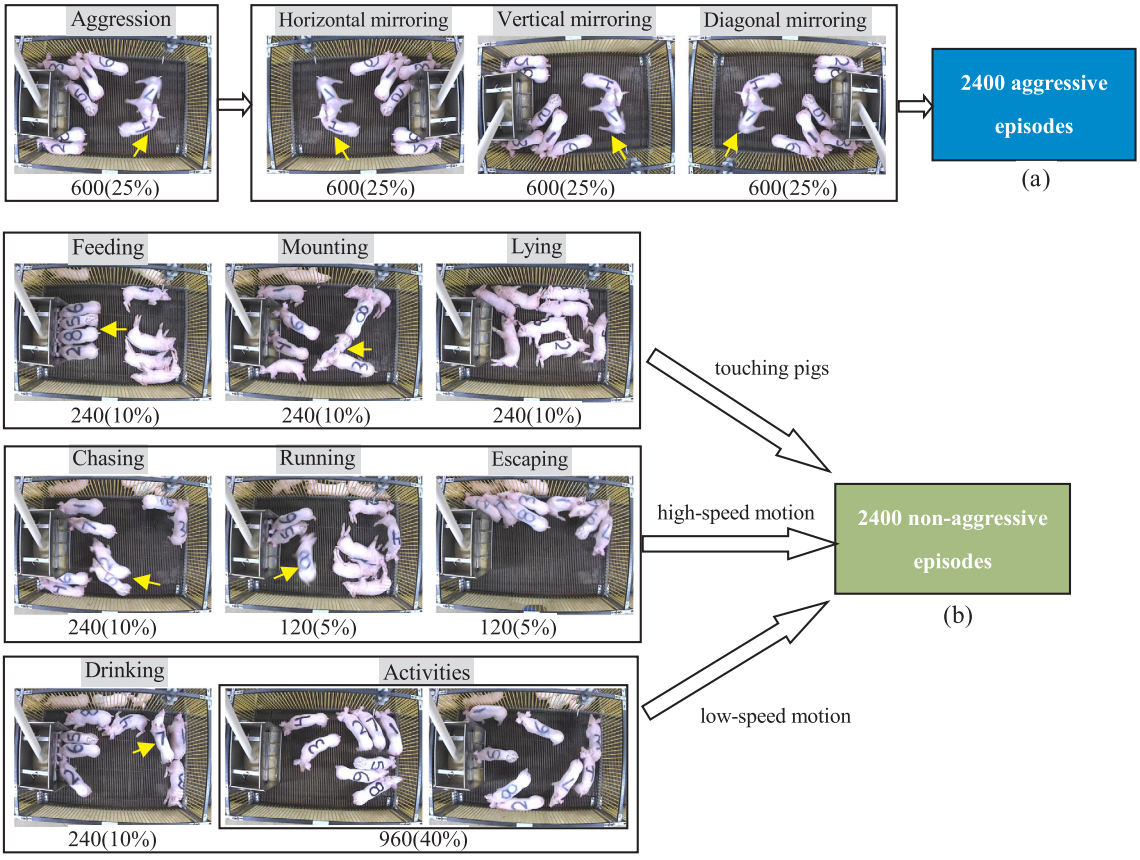

Fig. 1. Allocation of aggressive and non-aggressive episodes.

frames.

In this study, aggressive 2 s-episodes were labelled as the vector [1, 0] and non-aggressive 2 s-episodes as the vector [0, 1] by an computer scientist with expertise in labelling aggressive behaviour of pigs.

2.2. Algorithm

The idea of the proposed method is as followed:

1. Firstly, CNN was used to transform the original image (3211264(=224 × 224 × 64)-dimensional vector) into a feature with discrimination (25088(=7 × 7 × 512)-dimensional vector) by feature dimension reduction and optimisation.
2. Then, a special neural network consisting of LSTM, fully connected layer and the *Softmax* function was used to classify aggressive and non-aggressive episodes. The input of this special neural network is the above 25088-dimensional vector, and the output is a 2-dimensional vector, i.e. [1, 0] representing aggression and [0, 1] representing non-aggression.

Fig. 2(a) illustrates schematic diagram of using CNN and LSTM to recognise aggressive episodes. In this study, a vision model (VGGNet) (Simonyan and Zisserman, 2015) was used and pre-trained on the MS-COCO dataset (Lin et al., 2014). Among them, CNN used 5 blocks of convolutional and pooling layers of VGG-16. The input of CNN was the resized RGB image with a resolution of 224 × 224 pixels. Through these 5 blocks of convolutional and pooling layers, this image was flatten into a 25088-dimensional vector (Fig. 2(b)).

Fig. 2(c) illustrates the schematic diagram of LSTM. LSTM can be considered as a special neuron with 4 inputs and 1 output. Where z, z_i, z_o and z_f are the control signal of LSTM. These 4 signals are input into the input gate, output gate and forget gate in order to obtain the output y^t. Memory units c^t and h^t generated in this process are brought into next LSTM. It makes LSTM have a memory function (t = 1, 2,…, 60). The activation functions g of z is the *tanh* function within the interval [-1, 1]. The activation function f of z_i, z_o and z_f is the *Sigmoid* function within the interval [0, 1]. The activation functions h of memory cell is the *tanh* function within the interval [−1, 1]. Eq. (1) was used to calculate c^t, h^t and y^t.

$$\begin{aligned} c^t = c' &= g(z)f(z_i) + cf(z_f) \\ h^t &= h(c') \\ y^t &= a = h(c')f(\boldsymbol{z_o}) \end{aligned} \tag{1}$$

Fig. 2(d) describes the CNN and LSTM network for recognition of aggressive episodes from the vector view. Firstly, VGG-16 was used to convert each frame in the video episode into a 25088-dimensional vector. In the first frame, this 25088-dimensional vector [x_1, x_2,…, x_{25088}] was multiplied with weights to obtain the control signals z, z_i, z_o and z_f, and then the output y^1 and the memory units c^1 and h^1 were obtained through LSTM. In the second frame, the corresponding another 25088-dimensional vector [x_1, x_2,…, x_{100352}] was multiplied with weights to obtain the control signals z, z_i, z_o and z_f, and then the output y^2 and the memory units c^2 and h^2 were obtained through LSTM. Among them, the memory units c^1 and h^1 in the first frame were brought into the second LSTM to determine the memory units c^2 and h^2 in the second frame. By using this method in turn, the 60-dimensional vector [y^1, y^2,…, y^{60}] corresponding to these 60 frames were used as the total output of these 60 LSTM. This 60-dimensional vector was converted into a 2-dimensional vector through fully connected layer. Then, the *Softmax* function was used to convert all the elements of this 2-dimensional vector into the values within the interval (0, 1) and normalise these values (the sum of all values is 1). Finally, the class with the highest probability was selected as the predicted value 1 and another dimension as 0. Among them, the vector [1, 0] represents

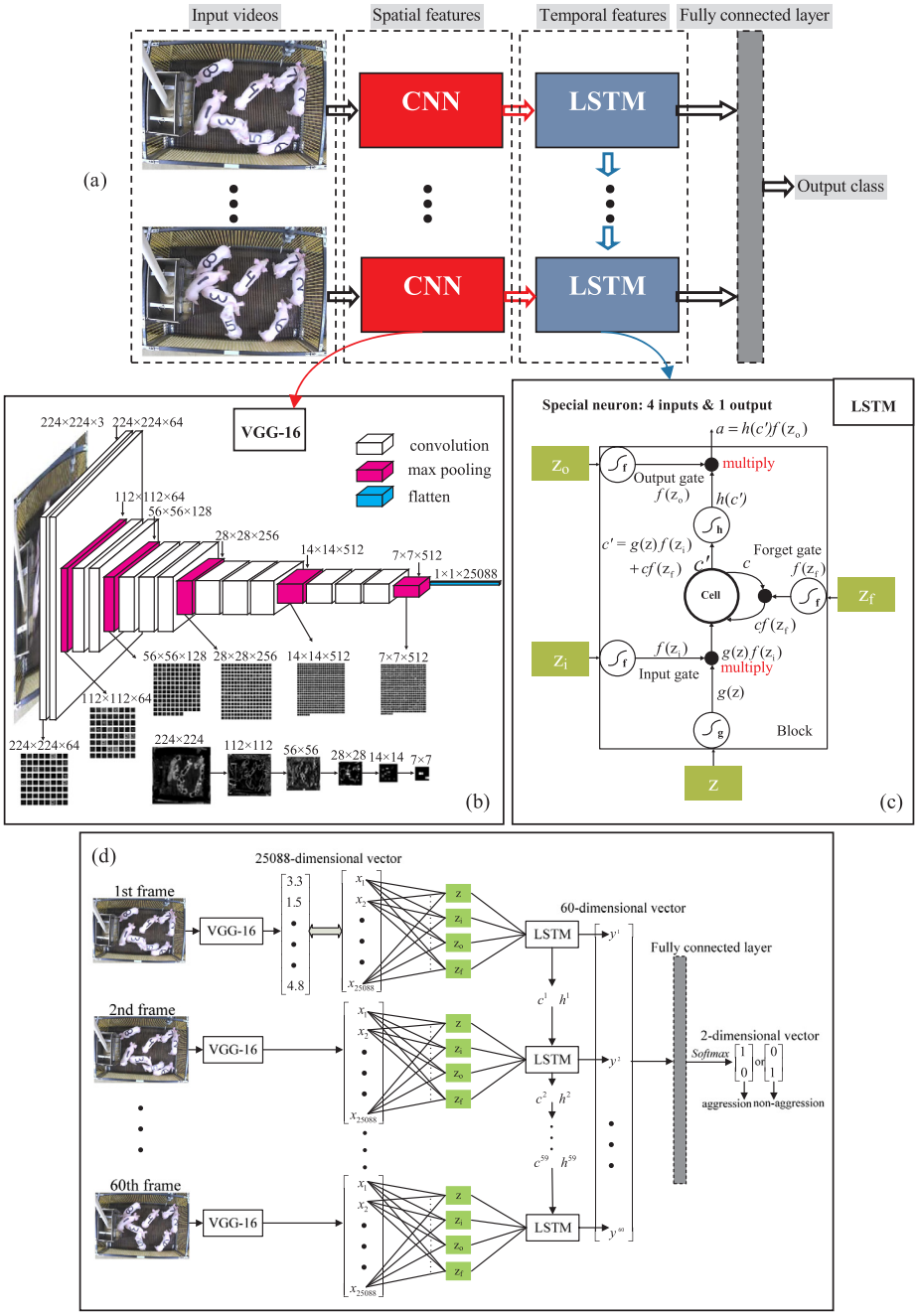

Fig. 2. Schematic diagram of using CNN and LSTM to recognise aggressive episodes.

aggression and the vector [0, 1] represents non-aggression.

In this paper, Eq. (2) was used to calculate the accuracy (ACC), false positive rate (FPR) and false negative rate (FNR) of the proposed method.

$$ACC = \frac{\text{Number of true positive and true negative episodes}}{\text{Total number of aggressive and non - aggressive episodes}} \times 100\%$$
$$FPR = \frac{\text{Number of false positive episodes}}{\text{Number of false positive and true negative episodes}} \times 100\%$$
$$FNR = \frac{\text{Number of false negative episodes}}{\text{Number of false negative and true positive episodes}} \times 100\% \quad (2)$$

where true positive episodes represent that aggressive episodes are correctly recognised as aggressive episodes. True negative episodes represent that non-aggressive episodes are correctly recognised as non-aggressive episodes. False positive episodes represent that non-aggressive episodes are falsely recognised as aggressive episodes. False negative episodes represent that aggressive episodes are falsely recognised as non-aggressive episodes.

Furthermore, cross-entropy function was used as the loss function in Eq. (3).

$$\text{loss} = -\sum_{c=1}^{M} y_c \log(p_c) \quad (3)$$

where c is the class, M is the number of all classes, y is the labelled result, and p is the predicted probability value normalised by the *Softmax* function. In this study, $M = 2$. Assuming that y is the vector [1, 0] representing aggression, and p is the vector [0.9, 0.1]. As a result, the calculation process of loss was shown in Eq. (4).

$$\text{loss} = -\sum_{c=1}^{2} y_c \log(p_c) = -y_1 \log(p_1) - y_2 \log(p_2) = -1 \log 0.9 - 0 \log 0.1$$
$$= 0.0457 \quad (4)$$

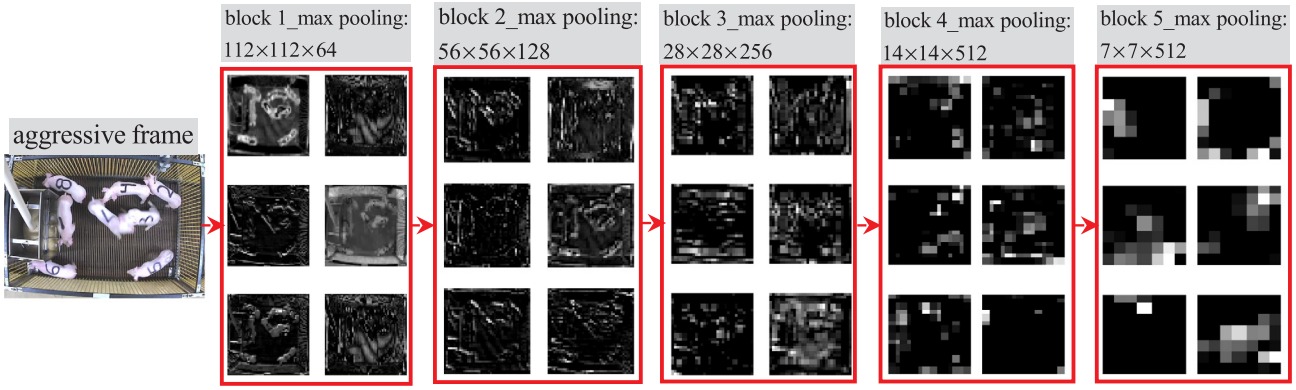

Fig. 3. Feature maps of max pooling layer of 5 blocks in VGG-16 corresponding to aggressive behaviours.

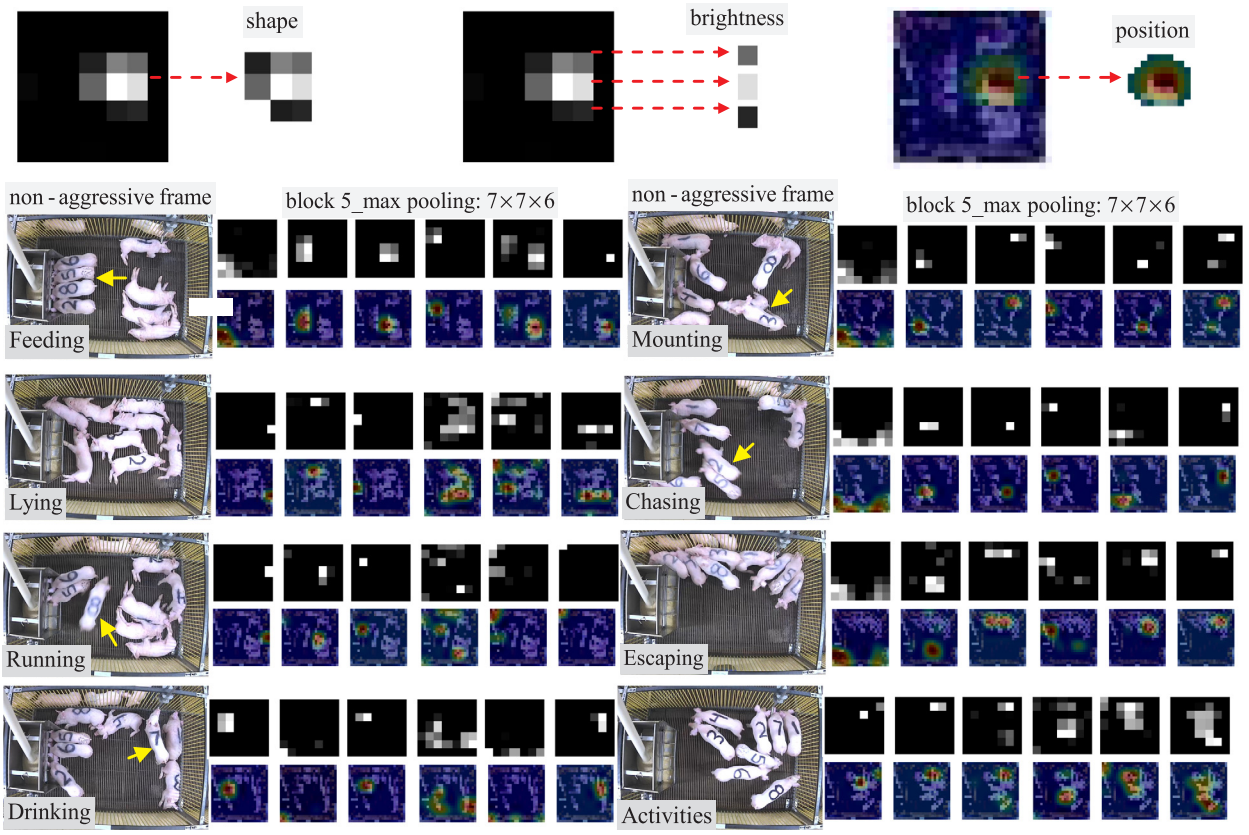

Fig. 4. Feature maps and heat maps of the last max pooling layer in VGG-16 corresponding to non-aggressive behaviours.

when p is the vector [0.9, 0.1], the final predicted result of the *Softmax* function is [1, 0], which is the same as the labelled result y. This result indicates that aggressive episodes are correctly recognised as aggressive episodes. Therefore, the number of true positive episodes will increase by 1.

3. Results and discussion

Fig. 3 illustrates the feature maps of max pooling layer of 5 blocks in VGG-16 corresponding to aggressive behaviours. From block 1 to block 5, 6-dimensional features were manually selected from the feature maps of pooling layer in each block for comparison. The red rectangles correspond to these 6-dimensional features. In block 1, 6-dimensional 112 × 112 features are similar to the original image. Through block 2 to block 5, the discrimination of features continuously increase from 6-dimensional 56 × 56 features to 6-dimensional 7 × 7 features. This discrimination is mainly reflected on shape and brightness. This result shows that the CNN feature of aggressive behaviours extracted by using VGG-16 has strong discrimination.

In order to further illustrate the discrimination of CNN features of non-aggressive behaviours, the heat map was adopted. Heat map can visualise which part of an image a CNN is looking at. In other words, the heat map is mainly used to represent the position used for extracting features in the feature map, and the hot pattern is used to visualise this position (Selvaraju et al., 2017). Therefore, the feature map mainly represents the shape and brightness information of the feature, and the heat map mainly represents the position information of the feature. Fig. 4 illustrates the feature maps and heat maps of the last max pooling layer in VGG-16 corresponding to non-aggressive behaviours. From 512-dimensional 7 × 7 feature maps and heat maps, the continuous 6-dimensional 7 × 7 features were manually selected for comparison. It can be seen that the 6-dimensional 7 × 7 features of different non-aggressive behaviours all have large differences in shape, brightness and position. This result shows that the CNN feature of non-aggressive

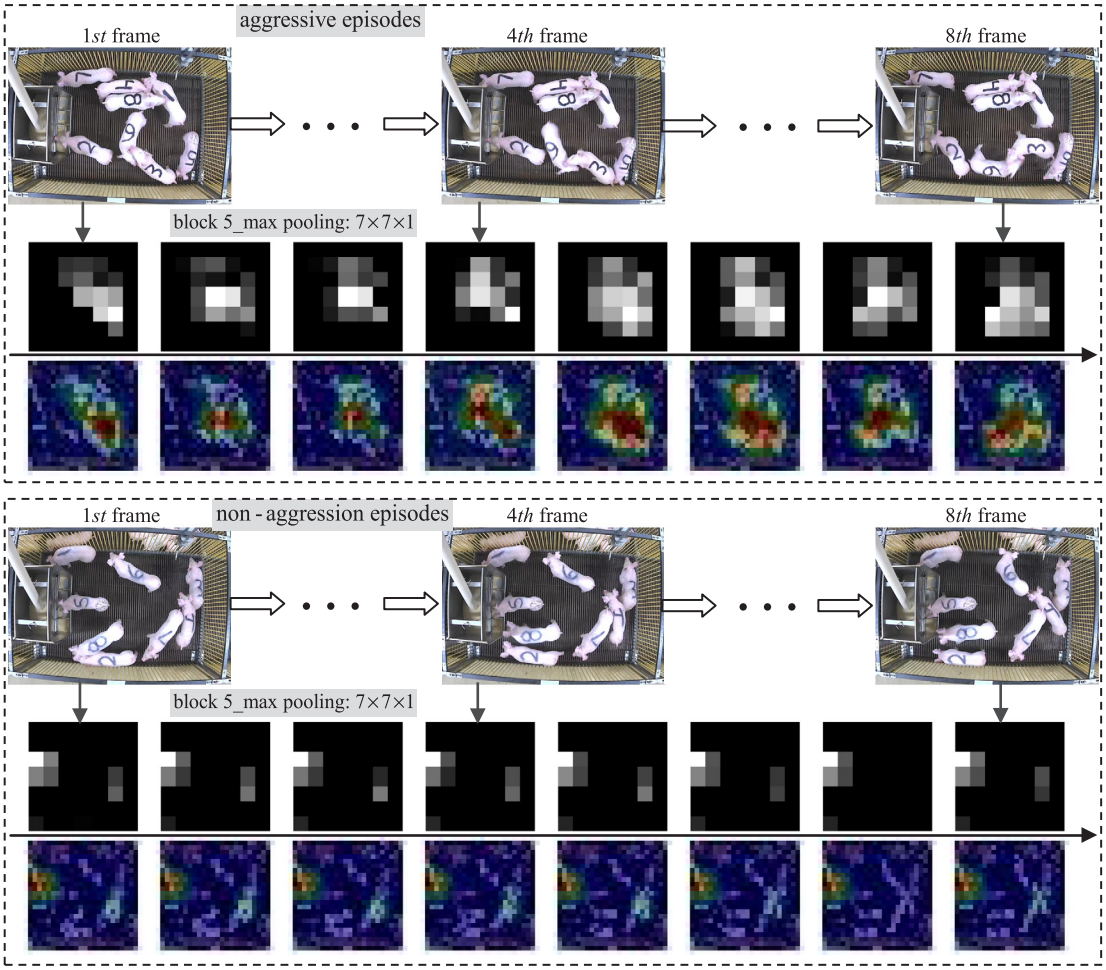

Fig. 5. Feature maps and heat maps of the last max pooling layer in VGG-16 corresponding to 8 continuous aggressive and non-aggressive frames.

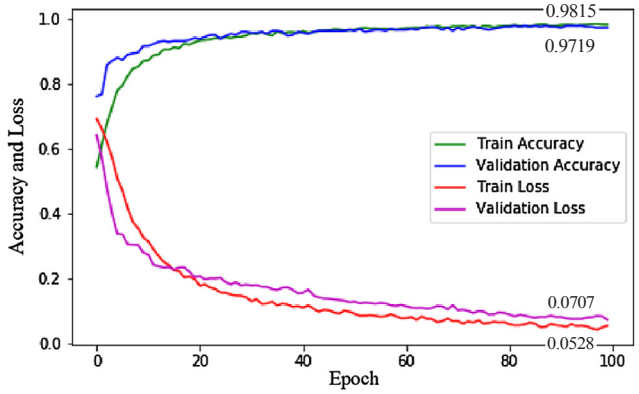

Fig. 6. Loss function and accuracy curves of the dataset of 2 s_60 frames.

behaviours extracted by using VGG-16 also has strong discrimination.

Fig. 5 illustrates the feature maps and heat maps of the last max pooling layer in VGG-16 corresponding to 8 continuous aggressive and non-aggressive frames. From the 512-dimensional 7 × 7 feature maps and heat maps of these 8 continuous frames, 1-dimensional 7 × 7 features of these 8 frames were manually selected for comparison. From the 1st frame to the 8th frame, it can be seen that the CNN feature of aggressive behaviours has a certain timing. This timing is mainly reflected in larger changes of shape, brightness and position. Moreover, from the 1st frame to the 8th frame, it can be seen that the CNN feature of non-aggressive behaviours also has a certain timing. However, this timing is mainly reflected in small changes of shape, brightness and position. These results indicate that CNN features can be input into LSTM frameworks to further extract temporal features. On the other hand, the CNN features between aggressive and non-aggressive behaviours have great difference in timing. This is the basis for using recurrent neural networks (i.e. LSTM and fully connected layers) to classify the CNN features of aggression and non-aggression. At the same time, this is consistent with the fast motion of aggression and the slow motion of non-aggression in the actual situation.

Fig. 6 illustrates the loss function and accuracy curves of the dataset of 2 s_60 frames. Where batchsize was set to 3. Through 100 epochs, the loss was reduced to 0.0707 and the accuracy was increased to 97.2%. The result indicates that combining VGG-16 and LSTM can be used to recognise aggressive episodes. While training and testing 4800 2 s-episodes (2.67 h) cost approximately 6.15 h. Assuming that 1 frame can be skipped, the total amount of data will be reduced by half, and the total time for training and testing will be also reduced by half. Therefore, this paper further verifies the feasibility of the proposed method under fewer frames and shorter time.

Fig. 7 illustrates the basis for improving the accuracy by skipping frames. When 1 frame was skipped, it can be seen that the discrimination of aggressive behaviours in aggressive sequence became larger. It is specifically reflected in the position and posture of the aggression, i.e. the position and size of the red rectangle (Fig. 7(a)). Meanwhile, when 1 frame was skipped, it can be seen that the discrimination of non-aggressive behaviours in non-aggressive sequence did not become larger (Fig. 7(b)). The results show that using the video episodes with skipping frames can increase the discrimination

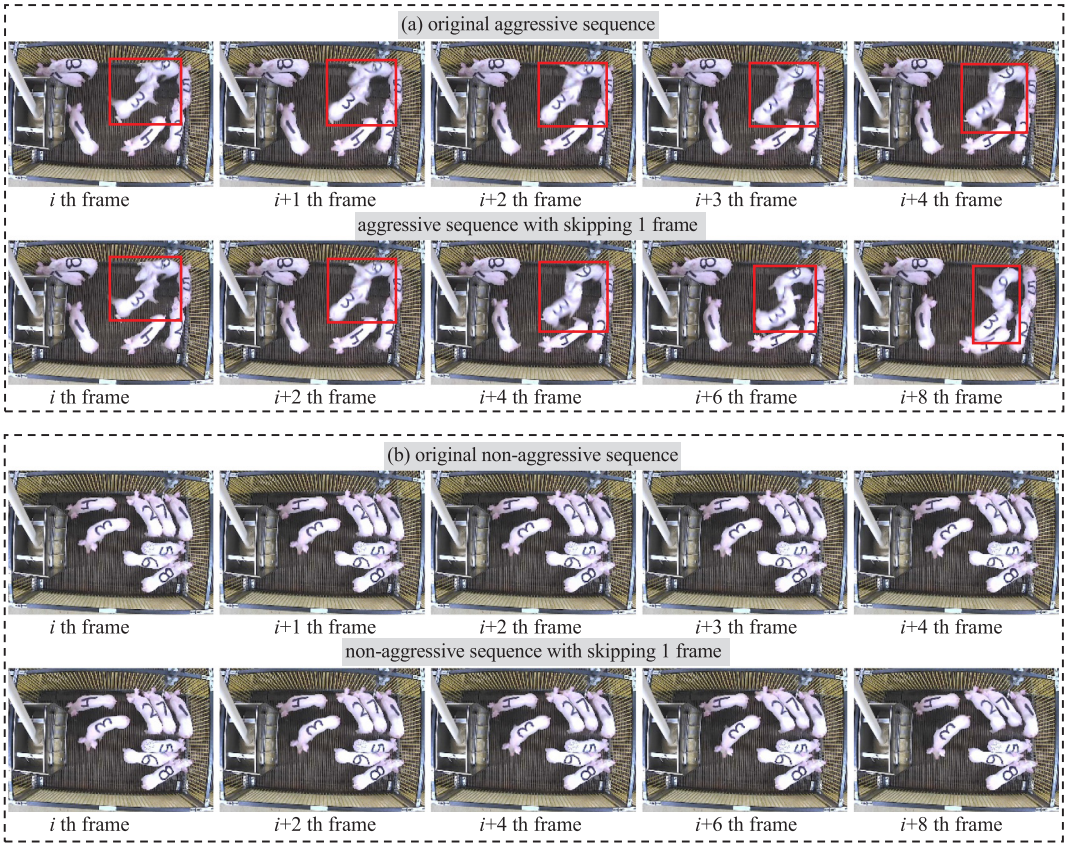

Fig. 7. Basis for improving the accuracy by skipping frames.

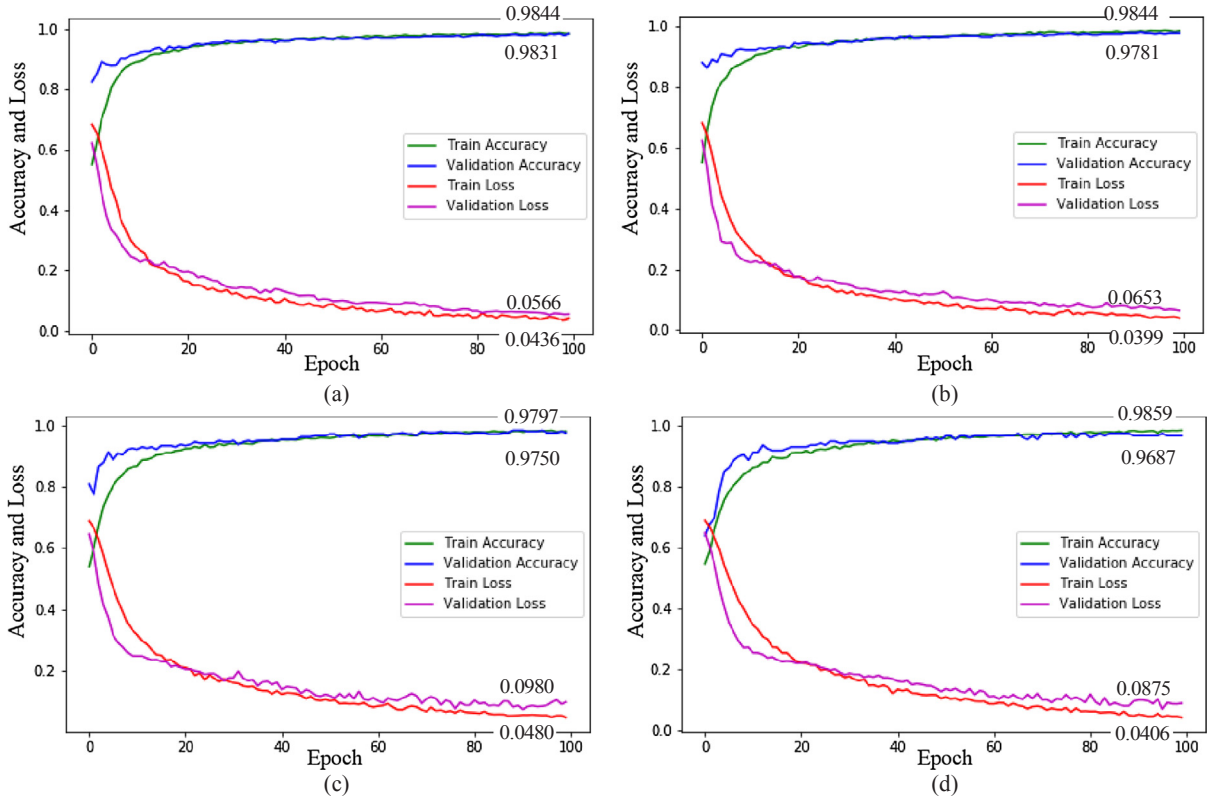

Fig. 8. Loss function and accuracy curves of the datasets under different number of skipped frames and different episode length: (a) 2 s_30 frames, (b) 2 s_20 frames, (c) 1 s_30 frames, and (d) 1 s_15 frames.

Table 1
Validation accuracy and validation loss when batchsize = 3, 5, 8, 10, 12, 16, 20, 24 and 32, respectively and epoch = 100 in the dataset of 2 s_30 frames.

Batechsize	3	5	8	10	12	16	20	24	32
Validation accuracy[%]	98.4	97.4	95.0	93.8	94.2	96.3	94.0	93.9	90.5
Validation loss	0.0566	0.1250	0.2400	0.2719	0.2943	0.2422	0.2534	0.2833	0.6476

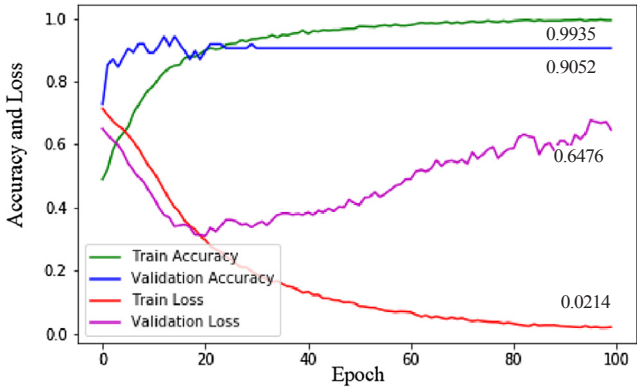

Fig. 9. Loss function and accuracy curves when batchsize = 32 in the dataset of 2 s_30 frames.

Table 2
The number of true positive (TP), false negative (FN), false positive (FP) and true negative (TN) episodes and the accuracy (ACC), false positive rate (FPR) and false negative rate (FNR) of recognising aggressive episodes of pigs.

TP	FN	FP	TN	ACC	FPR	FNR
2352	48	28	2372	98.4%	1.2%	2.0%

difference between aggressive and non-aggressive behaviours. This will make it possible to improve the accuracy of the proposed method. On the other hand, between adjacent frames, 2 aggressive pigs showed large changes in position and posture, while other non-aggressive pigs were close to stationary (Fig. 7(a)). As a result, non-aggressive pigs can be approximately considered as the background when aggression occurs. This is the reason why the CNN features of the entire image can be used to distinguish between aggression and non-aggression based on the entire pig herd rather than just on the 2 aggressive pigs.

Fig. 8 illustrates the loss function and accuracy curves of the datasets under different number of skipped frames and different episode length. Using the dataset of 2 s_30 frames, the accuracy was 98.4% with a loss of 0.0566. Using the dataset of 2 s_20 frames, the accuracy was 97.8% with a loss of 0.0653. Using the dataset of 1 s_30 frames, the accuracy was 97.5% with a loss of 0.0980. Furthermore, the accuracy was 96.9% with a loss of 0.0875 by using the dataset of 1 s_15 frames. Compared to Fig. 6, it can be seen that the accuracy was increased from 97.2% to 98.4% and the loss was reduced from 0.0707 to 0.0566 after changing 2 s_60 frames into 2 s_30 frames by skipping 1 frame. Moreover, the total time for training and test was decreased from 6.15 h to 3.22 h. These results indicate that setting the dataset as 2 s_30 frames can obtain the highest accuracy and the minimum loss. In the recognition of other behaviours, we can also record training set in high frame rate and perform this same analysis every time we train a model, then use the optimal frame rate for application of the trained model. For example, in the previous study of lying behaviour of pigs (Nasirahmadi et al., 2015), the original frame rate can be decreased as the postures of lying pigs are very similar in adjacent frames.

Table 1 illustrates the validation accuracy and validation loss when batchsize = 3, 5, 8, 10, 12, 16, 20, 24 and 32, respectively and epoch = 100 in the dataset of 2 s_30 frames. When batchsize was increased from 3 to 32, it can be seen that the accuracy was reduced from 98.4% into 90.5% and the loss was increased from 0.0566 to 0.6476. Therefore, the batchsize was set to 3 in this paper.

Fig. 9 illustrates the loss function and accuracy curves when batchsize was equal to 32 in the dataset of 2 s_30 frames. In Fig. 9, a significant overfitting phenomenon occured. Because when batchsize is too large and the total amount of data is constant, the number of units

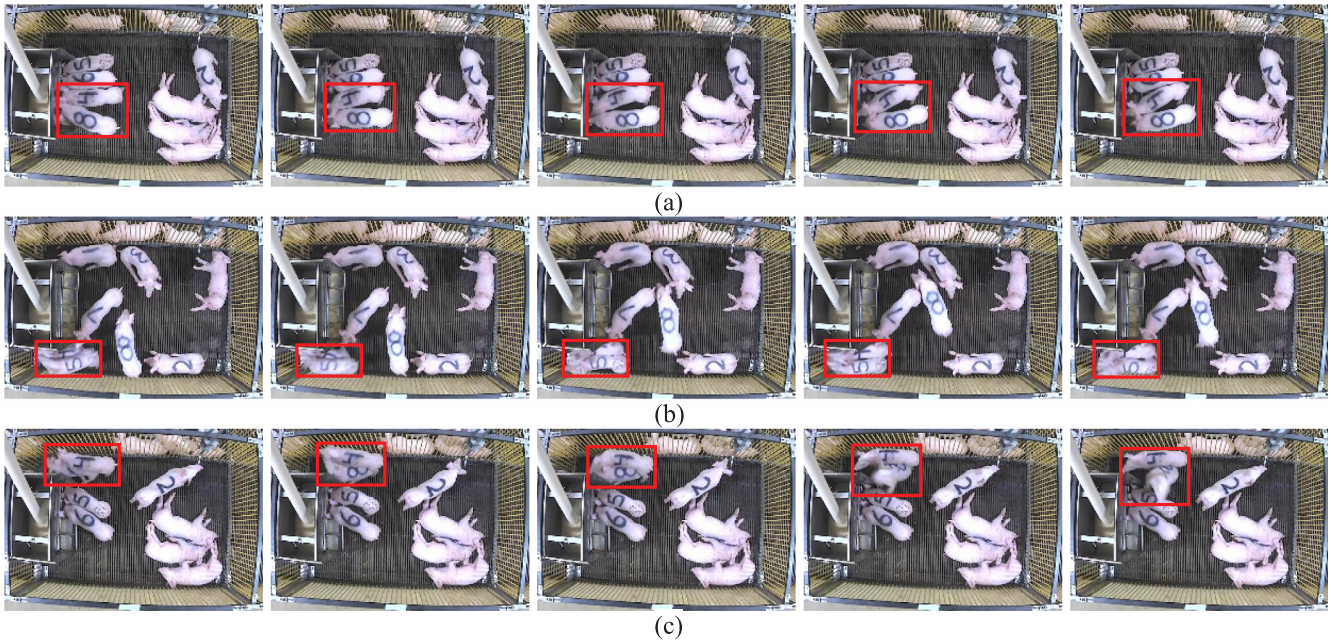

Fig. 10. Reasons for false recognition: (a-b) aggression falsely recognised as non-aggression, and (c) non-aggression falsely recognised as aggression.

used for training and testing in each epoch becomes smaller. As a result, the model does not learn effectively. The solution to this problem is to further increase the amount and diversity of the data.

Table 2 illustrates the number of true positive (TP), false negative (FN), false positive (FP) and true negative (TN) episodes and the accuracy (ACC), false positive rate (FPR) and false negative rate (FNR) of recognising aggressive episodes of pigs. It can be seen that the proposed method has good performance in recognition of aggressive episodes of pigs.

Fig. 10 illustrates the reasons for false recognition. False recognition is divided into 2 categories.

(1) Aggression is falsely recognised as non-aggression. In Fig. 10(a), pig 4 and pig 8 displayed a slight head to head knocking while feeding, which was falsely recognised as non-aggression. In Fig. 10(b), aggressive pig 4 and pig 5 showed parallel pressing at the corner between the feeding box and the pen, but they were limited by the narrow space and thus there was no larger displacement and shape change. Thus, this event was falsely recognised as non-aggression.
(2) Non-aggression is falsely recognised as aggression. In Fig. 10(c), during the mounting process of pig 8 and pig 4, the bottom pig 8 violently resisted and generated a larger displacement. Therefore, this event was falsely recognised as aggression.

Compared to previous studies (Chen et al., 2017; Chen et al., 2018), this work more completely distinguishes high-speed non-aggressive behaviours (i.e. running, chasing and escaping) from aggressive behaviours. These 3 behaviours and aggressive behaviours all involve high-speed motion, and thus it is difficult to distinguish among them by setting the thresholds of acceleration and kinetic energy. However, compared to aggression, these 3 non-aggressive behaviours have different spatial-temporal patterns. For example, running corresponds to continuous high-speed and non-adhesive motion of a single individual. Chasing corresponds to continuous high-speed and non-high-proportion adhesive motion of 2 individuals. Furthermore, escaping corresponds to continuous high-speed and high-proportion adhesive motion of the whole pig herd, while aggression corresponds to continuous high-speed and high-proportion adhesive motion of 2 individuals. In this paper, the spatial-temporal features extracted by using CNN and LSTM have a strong ability to discriminate among those behaviours. Thus, spatial-temporal features can be used to classify the above spatial-temporal patterns of the high-speed non-aggressive behaviours and aggressive behaviours.

Compared to previous work by Yang et al. (2018b) using deep learning on pig behaviours, the biggest characteristic of this work is combining CNN and LSTM to directly obtain spatial-temporal features by training videos rather than indirectly designing temporal features after obtaining spatial features by training frames. This makes the algorithm simpler and more straightforward in practice. Another characteristic of the proposed method is the use of whole group recognition rather than individual recognition (Chen et al., 2017; Chen et al., 2018), removing the step of locating 2 aggressive pigs within the group. In practical applications, the degree of crowding of pigs increases in commercial conditions, resulting in severe adhesion and overlapping among pigs. As a result, it is difficult to continuously extract 2 aggressive pigs from the herd. Therefore, the proposed method has more practical application value in a production environment.

In other previous studies of pig behaviours based on deep learning (Yang et al., 2018b; Zheng et al., 2018; Zhang et al., 2019), the position and class information of detected objects were obtained by using deep learning methods to train and test frames. In this paper, the spatial-temporal features of video episodes were directly extracted by combing CNN and LSTM, and thus the spatial position, class and temporal information were further obtained. Additionally, compared to previous studies of aggressive behaviours of pigs based on computer vision (Viazzi et al., 2014; Oczak et al., 2014; Lee et al., 2016; Chen et al., 2019), the proposed method solves the problem of low image quality and touching pig-bodies to a certain extent, by using VGG-16 to extract CNN features.

Many studies in the literatures support the idea that the aggressive behaviours are similar across all stages of a pig's life, with similar types of motion and interactions among pigs such as pigs biting each other's heads and shoulders as they rotate together (Jensen, 1994; Geverink et al., 1996; Oczak et al., 2013). Therefore, the method of recognising aggressive behaviours in pigs during the nursery period reported in this paper can be referenced when developing approaches to recognise aggressive behaviours of pigs in other life stages. However, to improve the robustness of the proposed method, it is still necessary to add aggressive episodes from pigs of different breeds, life stages and stocking densities as datasets in the future.

4. Conclusion

A deep learning method based on convolutional neural network and long short-term memory was used to recognise aggressive episodes of pigs with an accuracy of 97.2%. A frame skipping approach whereby 30 fps was reduced into 15 fps within each 2 s-episode improved the accuracy to 98.4% and reduced the total running time nearly in half. Using VGG-16 to extract CNN features had a stronger discrimination power for detecting aggressive behaviours. Using the CNN feature in an LSTM network can further extract temporal information. The cross entropy loss function and the accuracy can be used to effectively evaluate the proposed method. With the continuous increasing of batchsize, over fitting will occur and thus the amount and diversity of the data need to be increased. In the future, aggressive episodes of pigs in different breeds, life stages and stocking densities will be added as datasets in order to improve the robustness of the proposed method.

5. Author statement

The authors claim that none of the material in the paper has been published or is under consideration for publication elsewhere.

Declaration of Competing Interest

We declare that we have no financial and personal relationships with other people or organizations that can inappropriately influence our work, there is no professional or other personal interest of any nature or kind in any product, service and/or company that could be construed as influencing the position presented in, or the review of, the manuscript entitled.

Acknowledgements

This work was a part of the project funded by the "National Natural Science Foundation of China", China (grant number: 31872399), the "National Institute of Food and Agriculture", United States (grant number: 2017-67007-26176) and the "China Scholarship Council", China (File No. 201808320269).

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compag.2019.105166.

References

Büttner, K., Scheffler, K., Czycholl, I., Krieter, J., 2015. Network characteristics and development of social structure of agonistic behaviour in pigs across three repeated rehousing and mixing events. Appl. Animal Behav. Sci. 168, 24–30.
Banerjee, I., Ling, Y., Chen, M.C., Hasan, S.A., Langlotz, C.P., Moradzadeh, N., Chapman, B., Amrhein, T., Mong, D., Rubin, D.L., Farri, O., Lungren, M.P., 2019. Comparative effectiveness of convolutional neural network (CNN) and recurrent neural network

(RNN) architectures for radiology text report classification. Artif. Intell. Med. 97, 79–88.
Chen, C., Zhu, W.X., Ma, C.H., Guo, Y.Z., Huang, W.J., Ruan, C.Z., 2017. Image motion feature extraction for recognition of aggressive behaviours among group-housed pigs. Comput. Electron. Agric. 142, 380–387.
Chen, C., Zhu, W.X., Guo, Y.Z., Ma, C.H., Huang, W.J., Ruan, C.Z., 2018. A kinetic energy model based on machine vision for recognition of aggressive behaviours among group-housed pigs. Livestock Science 218, 70–78.
Chen, C., Zhu, W., Liu, D., Steibel, J., Siegford, J., Wurtz, K., Han, J., Norton, T., 2019. Detection of aggressive behaviours in pigs using a RealSence depth sensor. Comput. Electron. Agric. 166, 105003.
Donahue, J., Hendricks, L.A., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., Darrell, T., 2015. Long-term recurrent convolutional networks for visual recognition and description. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 2625–2634.
Erhard, H.W., Mendl, M., Ashley, D.D., 1997. Individual aggressiveness of pigs can be measured and used to reduce aggression after mixing. Appl. Animal Behav. Sci. 54 (2), 137–151.
Geverink, N.A., Engel, B., Lambooij, E., Wiegant, V.M., 1996. Observations on behaviour and skin damage of slaughter pigs and treatment during lairage. Appl. Animal Behav. Sci. 50 (1), 1–13.
Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural Comput. 9 (8), 1735–1780.
Jensen, P., 1994. Fighting between unacquainted pigs - effects of age and of individual reaction pattern. Appl. Animal Behav. Sci. 41 (1), 37–52.
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, European conference on computer vision, ECCV 2014: Computer Vision – ECCV 2014, 740–755.
Lee, J., Jin, L., Park, D., Chung, Y., 2016. Automatic recognition of aggressive behavior in pigs using a kinect depth sensor. Sensors 16, 631.
McGlone, J.J., 1985. A quantitative ethogram of aggressive and submissive behaviours in recently regrouped pigs. J. Anim. Sci. 61 (3), 559–565.
Nasirahmadi, A., Richter, U., Hensel, O., Edwards, S., Sturm, B., 2015. Using machine vision for investigation of changes in pig group lying patterns. Comput. Electron. Agric. 119, 184–190.
Nguyen, H.T., Nguyen, C.T., Bao, P.T., Nakagawa, M., 2018. A database of unconstrained Vietnamese online handwriting and recognition experiments by recurrent neural networks. Pattern Recogn. 78, 291–306.
O'Malley, C.I., Wurtz, K.E., Steibel, J.P., Bates, R.O., Ernst, C.W., Siegford, J.M., 2018. Relationships among aggressiveness, fearfulness and response to humans in finisher pigs. Appl. Animal Behav. Sci. 205, 194–201.
Oczak, M., Viazzi, S., Ismayilova, G., Sonoda, L.T., Roulston, N., Fels, M., Bahr, C., Hartung, J., Guarino, M., Berckmans, D., Vranken, E., 2014. Classification of aggressive behaviour in pigs by activity index and multilayer feed forward neural network. Biosyst. Eng. 119 (4), 89–97.
Oczak, M., Ismayilova, G., Costa, A., Viazzi, S., Sonoda, L.T., Fels, M., Bahr, C., Hartung, J., Guarino, M., Berckmans, M., Vranken, E., 2013. Analysis of aggressive behaviours of pigs by automatic video recordings. Comput. Electron. Agric. 99, 209–217.
Peden, R.S.E., Turner, S.P., Boyle, L.A., Camerlink, I., 2018. The translation of animal welfare research into practice: the case of mixing aggression between pigs. Appl. Animal Behav. Sci. 204, 1–9.
Spoolder, H.A.M., Edwards, S.A., Corning, S., 2000. Aggression among finishing pigs following mixing in kennelled and unkennelled accommodation. Livest. Product. Sci. 63 (2), 121–129.
Simonyan, K., Zisserman, A., 2015. Very deep convolutional networks for large-scale image recognition. ICLR.
Srivastava, N., Mansimov, E., Salakhutdinov, R., 2015. Unsupervised learning of video representations using LSTMs. International Conference on Machine Learning.
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, pp. 618–626 22-29 October.
Tsironi, E., Barros, P., Weber, C., Wermter, S., 2017. An analysis of convolutional long short-term memory recurrent neural networks for gesture recognition. Neurocomputing 268, 76–86.
Tian, M., Guo, H., Chen, H., Wang, Q., Long, C., Ma, Y., 2019. Automated pig counting using deep learning. Comput. Electron. Agric. 163, 104840.
Viazzi, S., Ismayilova, G., Oczak, M., Sonoda, L.T., Fels, M., Guarino, M., Vranken, E., Hartung, J., Bahr, C., Berckmans, D., 2014. Image feature extraction for classification of aggressive interactions among pigs. Comput. Electron. Agric. 104 (2), 57–62.
Yang, A., Huang, H., Zheng, C., Zhu, X., Yang, X., Chen, P., Xue, Y., 2018a. High-accuracy image segmentation for lactating sows using a fully convolutional network. Biosyst. Eng. 176, 36–47.
Yang, Q., Xiao, D., Lin, S., 2018b. Feeding behavior recognition for group-housed pigs with the Faster R-CNN. Comput. Electron. Agric. 155, 453–460.
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. Comput. Electron. Agric. 147, 51–63.
Zhang, Y., Cai, J., Xiao, D., Li, Z., Xiong, B., 2019. Real-time sow behavior detection based on deep learning. Comput. Electron. Agric. 163, 104884.