Computers and Electronics in Agriculture 193 (2022) 106650

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

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

# C3D-ConvLSTM based cow behaviour classification using video data for precision livestock farming

Yongliang Qiao [a,1], Yangyang Guo [b,*,1], Keping Yu [c], Dongjian He [b]

[a] Australian Centre for Field Robotics (ACFR), Faculty of Engineering, The University of Sydney, NSW 2006, Australia
[b] College of Mechanical and Electronic Engineering, Northwest A&F University, Yangling, Shaanxi 712100, China
[c] Global Information and Telecommunication Institute, Waseda University, Tokyo 169-0072, Japan

ARTICLE INFO

*Keywords:*
Behaviour classification
3D CNN
LSTM
Deep learning
Precision livestock farming

ABSTRACT

Cow behaviour provides valuable information about animal welfare, activities and livestock production. Therefore, monitoring of behaviour is gaining importance in the improvement of animal health, fertility and production yield. However, recognizing or classifying different behaviours with high accuracy is challenging, because of the high similarity of movements among these behaviours. In this study, we propose a deep learning framework to monitor and classify dairy behaviours, which is intelligently combined with C3D (Convolutional 3D) network and ConvLSTM (Convolutional Long Short-Term Memory) to classify the five common behaviours included feeding, exploring, grooming, walking and standing. For this, 3D CNN features were firstly extracted from video frames using C3D network; then ConvLSTM was applied to further extract spatio-temporal features, and the final obtained features were fed to a softmax layer for behaviour classification. The proposed approach using 30-frame video length achieved 90.32% and 86.67% classification accuracy on calf and cow datasets respectively, which outperformed the state-of-the-art methods including Inception-V3, SimpleRNN, LSTM, BiLSTM and C3D. Additionally, the influence of video length on behaviour classification was also investigated. It was found that increasing video sequence length to 30-frames enhanced classification performance. Extensive experiments show that combining C3D and ConvLSTM together can improve video-based behaviour classification accuracy noticeably using spatial–temporal features, which enables automated behaviour classification for precision livestock farming.

## 1. Introduction

Accurate automatic management based on video surveillance has gradually penetrated into many fields of livestock and poultry farming, among which the behaviour identification of breeding objects and the decision of livestock management schemes have become hot topics in current research (Meunier et al., 2018). Behaviour mainly refers to the activities as they respond to the environment or the way they express themselves, which is often a time series activity consistent and predictable (Moran and Doyle, 2015). In the livestock sector, behaviour is a valuable indicator of health, production and general well-being. The changing of physical behaviour including exploratory activity, reproductive activity, food and water intake, grooming and other social behaviours is often regarded as an early detector of diseases (e.g. lameness, injury, oestrus or pregnancy). Therefore, monitoring and classification of behaviour play an important role in timely management decisions to optimize animal performance, welfare and environmental outcomes in precision livestock farming (PLF) (Qiao et al., 2021).

However, the challenge in using sensor data is to automate the differentiation of behavioural activities. Consequently, there is a need to develop analytic models that make behaviour classification more accurate, robust and universally reusable.

Motion sensors (accelerometer, IMU, pedometer and magnetometer), Global Positioning Systems (GPS) and wireless sensor networks (WSNs) integrated systems have been designed to track and monitor animal behaviour over a large range in the PLF (Qiao et al., 2021). Achour et al. (2021) utilized an efficient-energy IMU sensor to continuously measure and classify the behavior of dairy cows housed in free stall. Arablouei et al. (2021) proposed cattle behavior approach using accelerometry data. Pavlovic et al. (2021) classified cattle behaviours

* Corresponding author.
*E-mail address:* gyy_113529@nwsuaf.edu.cn (Y. Guo).
[1] The Authors contributed equally and share first authorship.

https://doi.org/10.1016/j.compag.2021.106650
Received 9 August 2021; Received in revised form 19 December 2021; Accepted 21 December 2021
Available online 6 January 2022

using neck-mounted accelerometer-equipped collars, which achieved an classification accuracy with an overall $F_1$ score above 0.82. However, the challenge in using sensor data is to automate the differentiation of behavioural activities. Consequently, there is a need to develop analytic models that make behaviour classification more accurate, robust and universally reusable.

In recent years, vision based approaches as a non-contact method have been used in PLF tasks such as animal identification, body contour extraction (Qiao et al., 2019; Xue et al., 2021; Balch et al., 2006), and activity monitoring (Peng et al., 2019; Li et al., 2021). Since behaviour is a time series activity, visual features of movement over time can be extracted from video to represent behaviour patterns. Gu et al. (2017a) used minimum bounding box and contour mapping to identify behaviour and hoof or back characteristics, which could help large-scale farming efficiently. To make the most of the video contents, one should consider the visual aspects and also characterize the object appearances as well as the motion present within the data.

Research on CNN models for video processing has considered learning spatio-temporal filters over raw sequence data and learning of frame-to-frame representations which incorporate instantaneous optic flow or trajectory-based models aggregated over fixed windows or segments (Li et al., 2018b). Tran et al. (2015) proposed a 3D convolution framework to learn temporal features for human action recognition, which outperforms frame-based CNNs and other algorithms. Although the spatio-temporal feature extracted from video demonstrated ability in livestock behaviour recognition, some challenges such as complex scenes, different grow stages, occlusion and tiny movements have limited its performance in the real industrial applications.

In this work, a C3D-ConvLSTM based behaviour classification approach using video data was proposed. By leveraging the strengths of both 3D convolution and ConvLSTM to capture long-range spatial and temporal dynamics, the proposed approach can significantly improve the accuracy of behaviour classification. More specifically, at each time step (image frame), a set of 3D CNN features, which describe both the visual content and the motion information, are firstly extracted from video data by using C3D network. Based on the extracted 3D CNN features, the ConvLSTM layers are then applied to further explore the temporal relationship between video frames by using temporal evolution for 3D CNN features. Capturing such information proves to be useful for behaviour classification. As we will show later in the paper, by taking the advantage of both 3D CNN features and ConvLSTM temporal information, the proposed approach improves the behaviour classification accuracy, compared to other state-of-the-art methods (e.g. Inception-V3, SimpleRNN, LSTM, BiLSTM and C3D).

The major contributions of this work are: (1) Introducing a 3D convolution neural network based behaviour classification framework, that is capable of automatically learning spatio-temporal features; (2) We intelligently incorporate C3D and ConvLSTM networks to capture long-range spatial and temporal dynamics. C3D extracts 3D CNN features from frames which contain richer motion information whilst ConvLSTM is capable of exploring the temporal relationship between video frames. The combining of C3D and ConvLSTM effectively fuses spatial and temporal features to enhance behaviour classification accuracy; (3) We have performed a systematic assessment of the proposed framework on two different datasets. The experimental results show that the proposed approach using 30-frame video length achieved 90.32% and 86.67% classification accuracy on calf and cow datasets respectively, which outperformed the state-of-the-art methods such as Inception-V3, SimpleRNN, LSTM, BiLSTM and C3D; (4) The influence of video frame length on behaviour classification accuracy was investigated. It was found that increasing video sequence length to 30-frames enhanced behaviour classification performance.

The remainder of this paper is organized as follows: Section 2 reviews the recent literature on behaviour recognition and classification; Section 3 introduces the datasets used and the proposed approach; Section 4 presents the experiment setup and performance evaluation method; Behaviour classification results and analysis can be seen in Section 5; Section 6 discusses the application's specifications, and then conclusions are given in Section 7.

## 2. Related works

### 2.1. On-body sensor based behaviour classification

On-body sensors (e.g. accelerometers, magnetometers and GPS) are directly in contact with the animal body, and behaviour can be identified and classified by analysing the collected information (Andriamandroso et al., 2017). Tamura et al. (2019) classified a cow's eating, rumination and lying behaviour using a 3-axis neck-mounted accelerometer placed on the neck. Rahman et al. (2018) proposed behaviour (feeding, standing or ruminating) classification using data from collar, halter, and ear tag sensors. Weizheng et al. (2019) used triaxial acceleration sensors to automatically recognize feeding, ruminating and other behaviours of dairy cows, which achieved 96.1% and 97.5% specificity for feeding and ruminating behaviour recognition respectively. Riaboff et al. (2020) used accelerometers and machine learning algorithms to classify grazing, walking and ruminating behaviours.

In addition, Benaissa et al. (2019) collected data of the neck- and the leg-mounted accelerometers and used supervised machine learning techniques to realize dynamic behavioural classification. Smith et al. (2015) developed a multi-class behaviour model to classify six different behaviours for steers fitted with 3-axis accelerometer and pressure sensors. Peng et al. (2019) proposed a framework that combined long short-term memory and inertial measurement units to classify behaviours, which achieved above 80% classification accuracy for eight behaviour patterns.

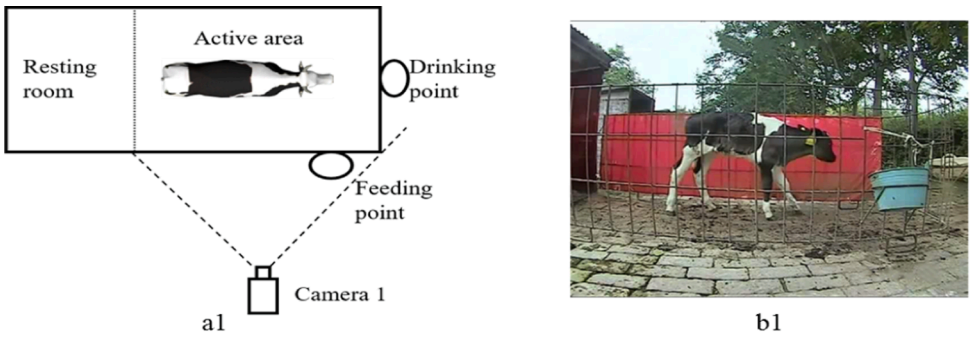
However the on-body sensor based approach is heavily reliant on the data quality and often influences the growth of cows. As these sensors directly contact the body, they may affect animal welfare or health over time. In addition, false readings would be read when they have bumped into fences or into other cows, or when cows are sick. Moreover, sensing devices are expensive for dairy farmers (Guo et al., 2019b).

### 2.2. Vision and machine learning based approaches

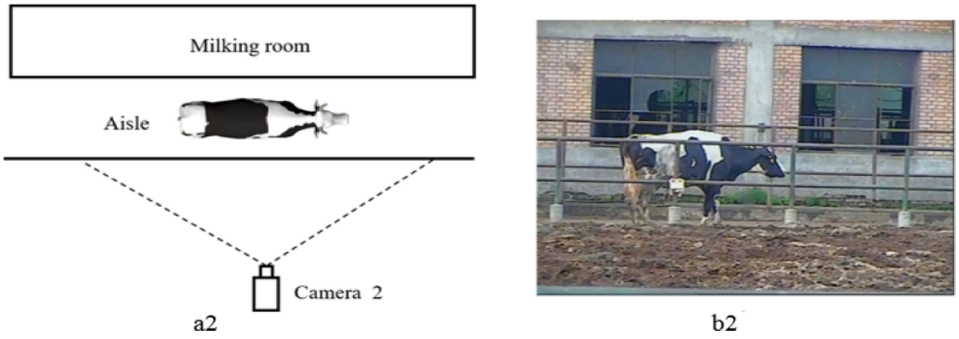
More recently, vision combing machine learning methods for continuous and non-intrusive behaviour monitoring and classification have attracted more attention in PLF (Liu et al., 2020). Ronghua et al. (2017) extracted cow object from video and obtained moving behaviour using dynamic tracking, then cow behaviour was recognized using dynamic analysis. Li et al. (2018a) developed an automatic cow monitoring system using an optical flow tracking algorithm and morphological features, which is helpful to analyze dairy cows' self-protective behaviors in the process of dairy cow breeding and management. Gu et al. (2017b) proposed an image analysis based cow reproduction and healthy behaviour identification method using surveillance video, and their experimental results showed that the recognition rates of hoof disease and heat in the reproduction and health of dairy cows are greater than 80%. Meunier et al. (2018) used image analysis to refine measurements of dairy cow behaviour from a real-time location system.

Besides, Guo et al. (2020) combined a machine vision-based method and scene information to identify the interaction behaviour of calf. Peng et al. (2020) developed an LSTM-RNN model to detect and recognize calving related behaviours, which achieved classification accuracy of feeding, ruminating (lying), ruminating (standing), lying normal, standing normal of 76.0%, 92.6%, 88.3%, 63.2%, 78.0% respectively. Salau and Krieter (2020) presented an automated camera-based system, which could be helpful to understand the herd behaviour by visualizing cows' usage of the barn.

However, most of the vision based approaches focus on the image features without paying much attention to temporal information.




(a) Calf data acquisition platform and its video effects




(b) Adult cow data acquisition platform and its video effects

**Fig. 1.** Video collection setup for calf and cow.

Moreover, the environment in a dairy farm is complex, image features are easy to be affected by the illumination or background. In addition, at different growth stages, cow behaviour classification results of the same model are also quite different. Therefore, a reliable method with high classification accuracy is demanded. In this paper, we propose a deep learning framework which intelligently combines C3D and ConvLSTM networks, and which exploits both time-domain information and spatial information to enhance behaviour classification of dairy calves and cows.

## 3. Material and methods

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

### 3.1. Data acquisition

In our experiment, both the calf and adult cow datasets were acquired from a commercial dairy farm in Yangling, China. In addition, due to the high mortality rate of newborn calves, calves within 6 months of age were raised separately in calf pens for fine feeding and management. Videos were recorded using a DS-2DM1-714 integrated IP camera (Hikvision Inc., Hangzhou, China), and the range of the focal length of the lens is 3.84–88.4 mm. Camera setup and video effects were shown in Fig. 1.

For calf data acquisition (Fig. 1 (a)), a two-month old Holstein dairy calf in a rectangular fenced enclosure (4 m× 2 m× 1.5 m) was selected, and video data were collected from 07:00 to 18:00 each day in July 2013. The camera on the length side of the fence monitored the calf activity from the side with a wide angle of view. The vertical height of the camera was half the height of the fence (i.e. 0.75 m), and the horizontal distance to the fence was set to cover the whole activity area of the calf.

**Table 1**
behaviour dataset information.

| Behaviour | Description | Image display |
|---|---|---|
| Feeding | Head is placed in feeding trough or over feeding trough while the cow is chewing | |
| Exploring | Head is in close proximity of or in contact with the ground | |
| Grooming | Head is turned towards abdomen groom the body with the tongue | |
| Walking | The position of body and four legs has moved | |
| Standing | The position of body and four legs is unchanged | |

In terms of adult cow data acquisition (Fig. 1 (b)), videos of Holstein cows in mid-lactation time were captured from 8:00 to 14:00 on sunny days in August 2013. The camera was positioned on the supported beam of a feeding shed at a height of 1.8 m and 35 m away from the alley. After milking, side-view images were acquired while cows were passing through a 2 m wide, 7 m long run with a concrete floor to a water trough.

In our experiment, over 40-h videos were acquired and recorded. The

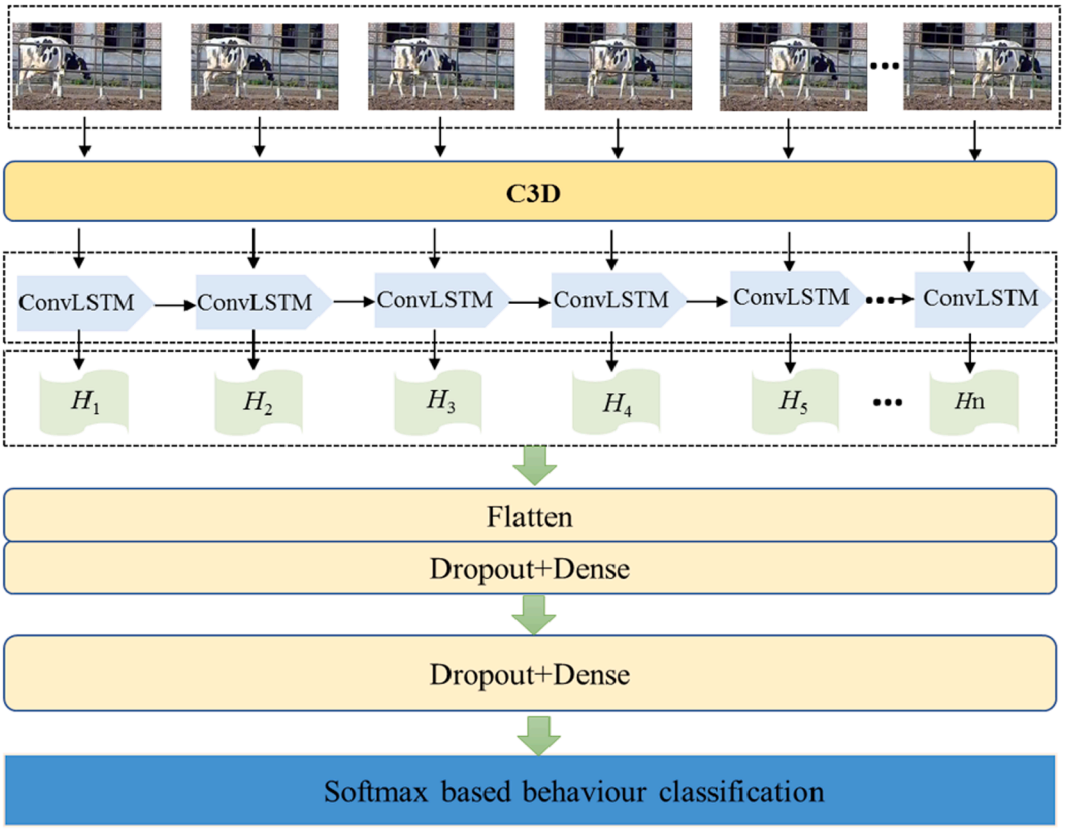



**Fig. 2.** The overall structure of C3D-ConvLSTM based behaviour classification approach.

videos contain behaviors, such as feeding, exploring, grooming, walking and standing. The behaviour description and statistics of calf video information obtained are shown in Table 1.

Both the calf and adult cow videos were captured with a frame rate of 25 fps and a code rate of 2000 kbps at a resolution of 704 × 576 pixels. Considered that there were some unnecessary areas (i.e. ground, wall and so on) in the original images, we extracted and saved the central part of the original image as Region of Interest (ROI). Thus the size of 704 × 477 and 704 × 291 images were obtained for calf and cow dataset respectively, after image ROI extraction.

### 3.2. C3D-ConvLSTM based behaviour classification

The overall structure of the proposed method is illustrated in Fig. 2, which uses C3D and ConvLSTM two key modules to classify behaviours from raw videos. The first module C3D is related to the extraction of 3D CNN features from videos, while the second module ConvLSTM is a kind of recurrent network which can capture long-term dynamics, and which preserves sequence information over time. Finally, the extracted features are fed into a soft-max layer for classifying different behaviours.

More specifically, for a given input cow video which contains $N$ frame images $\{I_i\}_{i=1}^{N}$, firstly, 3D CNN features of the video were extracted through C3D network. In C3D network, kernel filter is a 3D kernel, each Conv3D layer is followed by a Rectified Linear Unit (ReLU) layer and a 3D Max Pooling layer with a pool size of 2 × 2 × 2. These extracted 3D CNN features $X$ represent both the visual content and the spatial information of cow image in the video. Then the extracted 3D CNN features $X$ were given as an input to the ConvLSTM module for further exploited spatial–temporal features. In our work, one ConvLSTM layer with 512 sizes of the unit was used. The ConvLSTM keeps the same spatial size as the outputs of the 3D CNN features and just shrinks the temporal length to 1. Then after passing flatten, dropout (ratio is 0.5) and dense layers, the obtained spatial–temporal features were fed to a softmax layer to classify cow behaviours. The proposed approach is helpful to solve the problem of video-based behaviour classification because C3D-ConvLSTM efficiently captured spatial–temporal and cow motion information from the video.

#### 3.2.1. C3D module for 3D feature extraction

C3D networks can learn 3D CNN features of video data directly from image pixels, which can be trained on large-scale video dataset and model the appearance and motion information simultaneously (Ouyang et al., 2019). Compared with 2D convolutional networks, C3D extends the convolution along with the temporal directions, thus it can learn both discriminative visual features and temporal relationships from continuous RGB frames without any pre-processing step. The special trainable filters of C3D network capture the input data and automatically extract the spatial–temporal features for behaviour classification. Therefore, C3D was adopted to extract video features in our work.

We denote 3D Convolution and pooling kernel size as $d \times k \times k$, where $d$ is kernel temporal depth and $k$ is kernel spatial size. The calculation of 3D convolution, ReLU activation and 3D max pooling are as follows:

- 3D convolution: The 3D convolution is achieved by convolving a 3D kernel to the cube formed by stacking multiple contiguous frames together. By this construction, the feature maps in the convolution layer are connected to multiple contiguous frames in the previous layer, thereby capturing motion information. Formally, the value at position (x, y, z) on the $j$th feature map in the $i$th layer is given by:

$$v_{ij}^{xyz} = f\{b_{ij} + \sum_{k=1}^{m} \sum_{p=0}^{P_i-1} \sum_{q=0}^{Q_i-1} \sum_{r=0}^{R_i-1} w_{ijk}^{pqr} v_{(i-1)k}^{(x+p)(y+q)(z+r)}\} \quad (1)$$

where, $P_i$, $Q_i$ and $R_i$ are the size of the three-dimensional convolution kernel, m is the number of $i$-1 layer feature maps; $v_{(i-1)k}^{(x+p)(y+q)(z+r)}$ is the

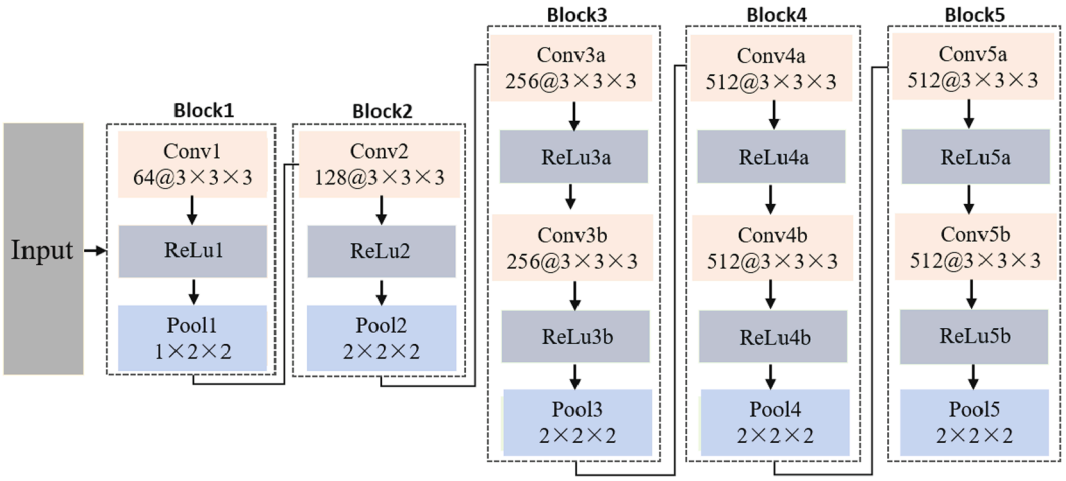



**Fig. 3.** The used C3D module for 3D CNN feature extraction.

value of the $k$-th feature map (x + p, y + q, z + r) in the $i$-1 layer; $w_{ijk}^{pqr}$ is the convolution kernel connected to the $k$-th feature map of the $i$-1 layer; $b_{ij}$ is partial set; $f(\cdot)$ is the ReLU activation function.

- ReLU activation function: ReLU is responsible for transforming the summed weighted input, which is a piecewise linear function that will output the input directly if it is positive, otherwise, it will output zero. The nonlinear function ReLU is used to introduce non-linearity in the ConvNet.
- 3D max pooling: The 3D pooling operation makes the feature cube in time dimension have a certain invariance, while greatly reducing the amount of calculation, thereby improving the three-dimensional convolutional neural network in the time dimension robustness. The formula for maximum pooling is as follows:

$$v_{pool}^{xyz} = \max_{0\leqslant i\leqslant S_1, 0\leqslant j\leqslant S_2, 0\leqslant k\leqslant S_3} \left(u_{x\times s+i, y\times t+j, z\times r+k}\right) \tag{2}$$

where, $u$ is the three-dimensional input vector, v is the output after the three-dimensional pooling operation, (s, t, r) is upsampling step size of coordinate (x, y, z) along x, y, z directions; $S_1 \times S_2 \times S_3$ is the sampling area size.

As illustrated in Fig. 3, the used C3D module has five blocks, the first two blocks have a 3D convolution layer, a ReLU activation function, and a 3D pooling layer, respectively. From block3 to block5, each block has two 3D convolution layers, two ReLU activation functions and one 3D pooling layer. The number of channels (filters) for 5 convolution layers from 1 to 5 is 64, 128, 256, 512, and 512 respectively. In our work, the $3 \times 3 \times 3$ small kernel sizes with stride $1 \times 1 \times 1$ is used in the convolution process, which is helpful to capture all changes in terms of spatial and temporal information. Intuitively, these different layers describe the visual content at different levels, each of which is complementary to each other for the task of behaviour classification.

In terms of the pooling layer, 3D maximum pooling was used. The pool1 layer, has a kernel size of $1 \times 2 \times 2$, with the goal of not merging the temporal signal and preserving the temporal information in the early phases. From pool2 to pool5 layers, $2 \times 2 \times 2$ pooling kernels with stride $2 \times 2 \times 2$ were used, which reduce the 3D CNN features by a factor of 8. Actually, all pooling layers keep the same number of feature maps as convolution layers with smaller spatial resolution.

By the multiple layers of convolution and subsampling, the input frames have been converted into a feature vector capturing the motion information from the cow videos. In the network, the output of each convolutional layer can be regarded as spatio-temporal features. The lower layers contain underlying features such as edges and colors whilst the higher layer outputs more semantic and discriminative features. In the C3D network, the last Conv5 layer has larger receptive fields and obtained the most invariant and discriminative features (complex visual elements).

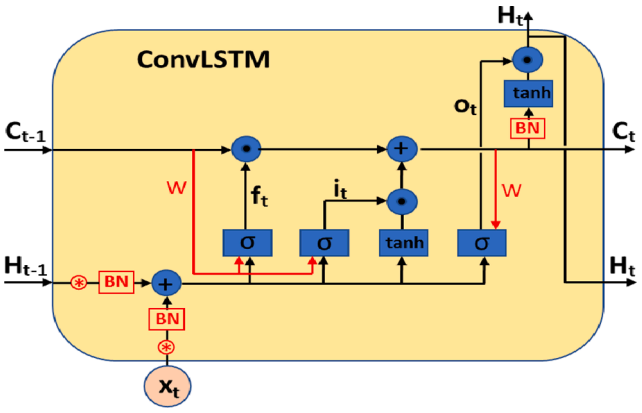



**Fig. 4.** The structure of ConvLSTM.

### 3.2.2. *ConvLSTM module*

To make the most of the video contents and learn complex temporal dynamics, ConvLSTM module was deployed to further extract spatio-temporal features by mapping input 3D CNN features to a sequence of hidden states. ConvLSTM is developed from Long Short Term Memory (LSTM), which performs better in sequential data problems (Wu et al., 2020).

The LSTM was prosed by extending RNN (Recurrent Neural Network) with memory cells. These cells usually have few linear interactions making the information maintaining process easier (Ordóñez and Roggen, 2016). Although LSTM can handle time-series data, for spatial data, it will bring redundancy, mainly due to the dependence of internal gates and similar feed-forward neural network calculations; it is impossible to characterize the local features of spatial data.

In order to exhibit dynamic temporal behavior, ConvLSTM was used in our work, which replaces feedforward calculations with convolution operations during the transition from input to state and state to state. Thus spatio-temporal relationships in the cow video data can be well modeled. As illustrated in Fig. 4, ConvLSTM consists of an input gate $i_t$, an forget gate $f_t$, an output gate $o_t$, and a memory cell $C_t$. Each gate is a nonlinear summation unit which controls the operation of the cell memory such as write (input gate), read (output gate) or reset (forget gate) (Itakura et al., 2019).

For a given input $X_t$ at time step $t$, through exploiting convolution operations into input-to-state and state-to-state transitions, the

**Table 2**
Description of the datasets employed in the experiments.

| Dataset | No.videos | Description |
|---|---|---|
| Calf | Train: 381 (102 feeding, 94 exploring, 60 grooming, 38 walking, 87 standing) Test: 93 (25 feeding, 23 exploring, 15 grooming, 9 walking, 21 standing) | major variations in illumination |
| Adult cow | Train:176 (24 exploring, 14 grooming, 91 walking, 47 standing) Test: 60 (8 exploring, 5 grooming, 31 walking, 16 standing) | major variations in illumination |

ConvLSTM captures underlying spatial features over long sequences of input data. The key equations of ConvLSTM can be formulated as follows:

$$i_t = \sigma(W_{xi} * X_t + W_{hi} * H_{t-1} + W_{ci} * C_{t-1} + b_i) \tag{3}$$

$$f_t = \sigma(W_{xf} * X_t + W_{hf} * H_{t-1} + W_{cf} * C_{t-1} + b_f) \tag{4}$$

$$C_t = f_t \odot C_{t-1} + i_t \odot tanh(W_{xc} * X_t + W_{hc} * {}_{t-1} + b_c) \tag{5}$$

$$o_t = \sigma(W_{xo} * X_t + W_{ho} * H_{t-1} + W_{co} * C_{t-1} + b_o) \tag{6}$$

$$H_t = o_t \odot tanh(C_t) \tag{7}$$

where, * denotes the convolution operator and $\odot$ denotes the Hadamard product; $\sigma$ is Sigmoid function; $W_x$ and $W_h$ are two-dimensional convolution kernel. The inputs $[X_1, \cdots X_t]$, cell outputs $[C_1, \cdots, C_t]$, hidden states $[H_1, \cdots, H_t]$, and gates $[i_t, f_t, o_t]$ are 3D tensors.

In this work, one layer of ConvLSTM was utilized, the ConvLSTM output was treated as a long-term space–time feature for cow behaviour distinguishing. The size of the convolution kernel was set as 5, the step size was set as 1, the number of filters was 512, padding mode was 'valid', which is the same as was used in the convolution operation. In addition, to avoid over-fitting during the training process, $L_2$ kernel regularizer and recurrent dropout were added also with value of 0.5.

### 3.3. Behavior classification using C3D-ConvLSTM

The spatio-temporal features generated by ConvLSTM layer represents the cow-behaviours in each cow video. Through a flatten layer, and dense layer with 0.5 dropout rate, these spatio-temporal features were fed to a softmax classifier for the behaviour classification. For a video is input to the C3D-ConvLSTM model, the output $y_t$ and probability value $P$ of the $n$-type behavior are obtained by the softmax layer:

$$P_m = \frac{e^{y_t^m}}{\sum_{n=1}^{n} e^{y_t^m}} \tag{8}$$

where, $m$ is the category label, the probability value of $P_m$ is in the range of 0 and 1. Index $m$ corresponding to the largest probability value $P_m$ is the final behaviour classification result. If it matches the ground truth class label, then it will be regarded as a true result. Otherwise is a false classification.

## 4. Experiment setup

### 4.1. Experimental dataset construction

To validate the effectiveness of the proposed C3D-ConvLSTM based cow behaviour classification, experiments were conducted on two different age datasets (i.e. calf and adult cow). In our work, the original long surveillance videos were cropped into 40-frame clips and there was no overlapping between two consecutive clips. Each clip is passed into our network to extract features. The details about training and testing datasets are presented in Table 2.

1) Calf dataset: 381 videos (i.e. 102 feeding, 94 exploring, 60 grooming, 38 walking, 87 standing) were used to train and 93 videos (i.e. 25 feeding, 23 exploring, 15 grooming, 9 walking, 21 standing) were used to test.
2) Adult cow dataset: 176 videos (24 exploring, 14 grooming, 91 walking, 47 standing) and 60 videos (8 exploring, 5 grooming, 31 walking,16 standing) were used to train and test respectively.

### 4.2. Comparison with state-of-the-art methods

To verify the effectiveness of the proposed approach, the proposed C3D-ConvLSTM based cow behaviour classification approach was compared with five state-of-the-art methods Inception-V3, Simple Recurrent Neural Network (SimpleRNN), LSTM, BiLSTM (Bidirectional LSTM) and C3D. Among these methods, Inception-V3 is a frame-based method, while the others are sequence/video-based approaches.

- Inception-V3 is one popular CNN model for image classification (Szegedy et al., 2016). In our experiment, images from training videos were used to train the network whilst the images from testing videos were regarded as the test dataset.
- SimpleRNN is one kind of RNN, which calculates hidden vector sequences and output vector sequences through a linear transform and an activation function (Guo et al., 2019a). In our work, one-layer SimpleRNN with 2048 cells was used.
- LSTM uses memory cells to maintain long-term information over long sequences of input data (Karim et al., 2019). One-layer LSTM (2048 cells) with a dropout rate of 0.5 are used in our work.
- BiLSTM consists of two independent LSTMs, which can sum up information from forward and backward directions of a sequence, and merge the information coming from the two directions (Li et al., 2020b). In our experiments, one BiLSTM layer with 2048 cells and 0.5 dropout rate was used.
- C3D network has 8 convolution layers and 5 pooling layers (each convolution layer is immediately followed by a pooling layer), 2 fully connected layers and a softmax layer.

### 4.3. Network training parameters

In our work, the proposed approach and the other methods are implemented by using Keras (Chollet, 2018) on a DELL TOWER PC with GeForce GTX 1080Ti GPU. In order to compare these methods fairly, all the network's input sizes were set to $112 \times 112 \times 3$. For the network training on each dataset, the training epoch was set to 1000, batch size was 10, and the learning rate was 0.00001. In addition, in order to avoid the influence of pre-trained weights, all the network's initial weights were random. The other parameters of each network were their default settings.

### 4.4. Performance evaluation

In order to evaluate performance, four widely used measures such as accuracy, precision, recall and $F_1$-score were used to evaluate the performance of behaviour classification. All the above four measures are ranged from 0% to 100%, high value means good predictive ability of the model. In addition, the results of testing were also arranged in confusion matrices, including true positive (tp), true negative (tn), false positive (fp), and false negative (fn). Accuracy, precision, recall and $F_1$ are defined, respectively, as follows:

$$Accuracy = \frac{tp + tn}{tp + tn + fp + fn} \times 100\% \tag{9}$$

$$Precision = \frac{tp}{tp + fp} \times 100\% \tag{10}$$

**Table 3**
Calf behaviour classification results (%).

| Methods | Accuracy | Precision | Recall | $F_1$ |
|---|---|---|---|---|
| Inception-V3 (frame) | 56.33 | ‡ | ‡ | ‡ |
| SimpleRNN | 74.19 | 83.94 | 66.08 | 73.95 |
| LSTM | 75.27 | 59.61 | 63.47 | 61.48 |
| BilSTM | 78.49 | 78.85 | 72.81 | 75.71 |
| C3D | 84.95 | 83.76 | 83.85 | 83.81 |
| C3D-ConvLSTM | **90.32** | 90.62 | 88.76 | 89.68 |

Note: ‡ means the value is below 50%.

$$Recall = \frac{tp}{tp + fn} \times 100\% \tag{11}$$

$$F_1 = 2 \times \frac{precision \times recall}{precision + recall} \times 100\% \tag{12}$$

## 5. Behaviour classification results

### 5.1. Calf behaviour classification results

The behaviour classification accuracy of C3D-ConvLSTM based approaches for calf dataset are illustrated in Table 3. The proposed C3D-ConvLSTM based approach achieved an accuracy of 90.32%, a precision of 90.62%, a recall of 88.76%, and an $F_1$ score of 89.68%. These yielded values are slightly higher than those of Inception-V3 (56.33% accuracy), SimpleRNN (74.19% accuracy), LSTM (75.27% accuracy), BiLSTM (78.49% accuracy) and C3D (84.95% accuracy). It also can be noticed that the $F_1$of the proposed C3D-ConvLSTM based approach is 89.68%, which is almost 5% higher than that of C3D (83.81% $F_1$). Experimental results illustrates that the proposed C3D-ConvLSTM could enhance the behaviour classification ability through extracting of 3D and temporal features.

The confusion matrix of the calf behaviour classification is displayed in Fig. 5. The proposed C3D-ConvLSTM achieved 100%, 53.33%, 90.48% and 100% classification accuracy for exploring, grooming, standing and feeding behaviours respectively, which are higher than that of C3D (95.65%, 46.67%, 80.95% and 96%). The proposed method obtained a high classification accuracy for calf's exploring, walking and feeding behaviours due to those features are significantly different from other behaviours. On the other hand, the classification accuracy of grooming is lowest, and almost half of the grooming behavior is wrongly classified as standing. The main reason is that grooming refers to head motion (towards the coat) based on standing, which is easy to be regarded as standing when calf head movement is slight.

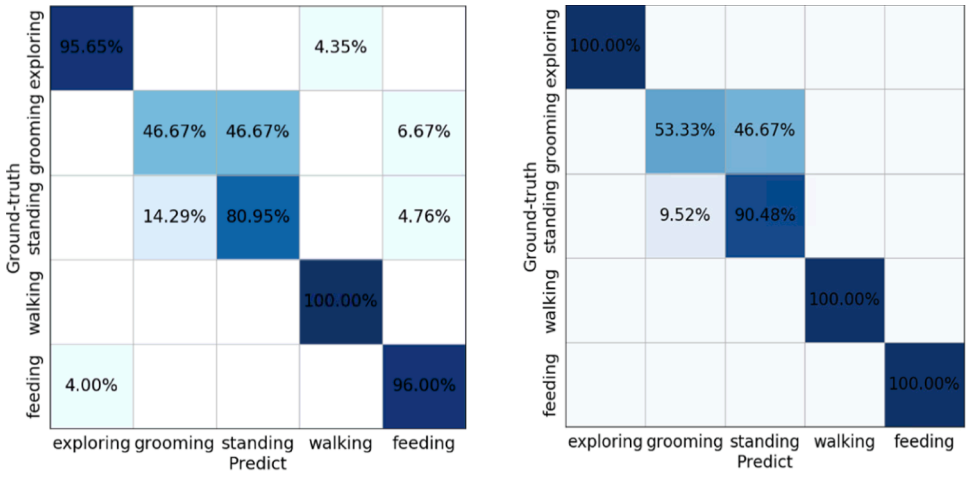
In addition, some true and false classification examples of the proposed C3D-ConvLSTM are shown in Fig. 6. In Fig. 6(a), calf's exploring, walking and feeding behaviours are well classified. However, it also can be seen in Fig. 6 (b) that some behaviours were misclassified due to the high similarity of movements among these behaviours. In addition, as there is a transition time between behaviour changing, the behaviours within the transition process had no obvious classification boundary resulting misclassification (Fig. 6 (b) A, B). Sometimes the misclassification of these two behaviours is caused by the background and lighting leading to the calf's head information being missed (Fig. 6 (b) C, F). In addition, the changing of calf postures in the scene (Fig. 6 (b) D, E) also reduce the classification accuracy.

### 5.2. Cow behaviour classification results

The behaviour classification results obtained by different methods is shown in Table 4. The proposed C3D-ConvLSTM based approach achieved an accuracy of 86.67%, a precision of 86.13%, a recall of 74.06%, and an $F_1$ score of 79.64% which is higher than those of comparison methods (i.e. Inception-V3, SimpleRNN, LSTM, BiLSTM and C3D). Specifically, the precision of the proposed C3D-ConvLSTM is almost 3% higher than that of C3D. All these results illustrate again that the proposed C3D-ConvLSTM could mine the spatial–temporal features of animal behaviour through 3D and temporal feature extraction, enhancing the behaviour classification ability.

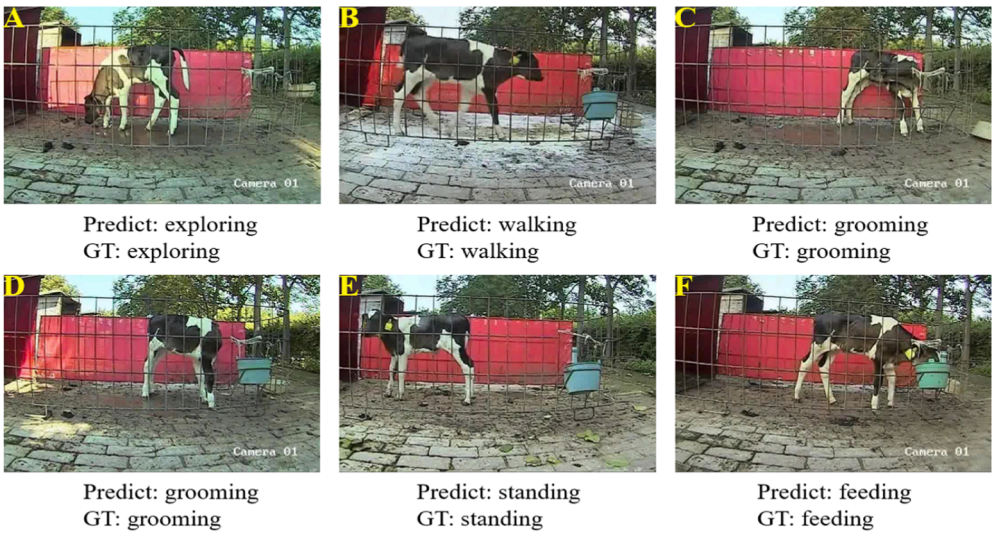
The confusion matrix of cow behaviour classification can be seen in Fig. 7. Both the C3D-ConvLSTM and C3D methods achieved 75% and 81.25% accuracy for the exploring and standing behaviours. But the classification accuracy of grooming and walking in C3D-ConvLSTM is slightly higher than that of C3D. The experimental results illustrate again that the proposed C3D-ConvLSTM has a strong feature representation ability for behaviour classification which outperformed comparison methods.

Some true and false classification examples of the proposed C3D-ConvLSTM are shown in Fig. 8. In Fig. 8 (b) A, some exploring behaviours were classified as walking behaviours due to lamed cow walking slowly with low head position. When the cow was exploring and the cow's head was occluded, this behaviour would be misclassified (Fig. 8 (b) B). In addition, cows would present different postures in the scene or the head turned to the other side, leading to misclassification of behaviour (Fig. 8 (b) C, D, F). And when there is more than one cow in the field of view, it can also cause misclassification (Fig. 8 (b) E). In general, the position of cow head, posture and transitional behaviours

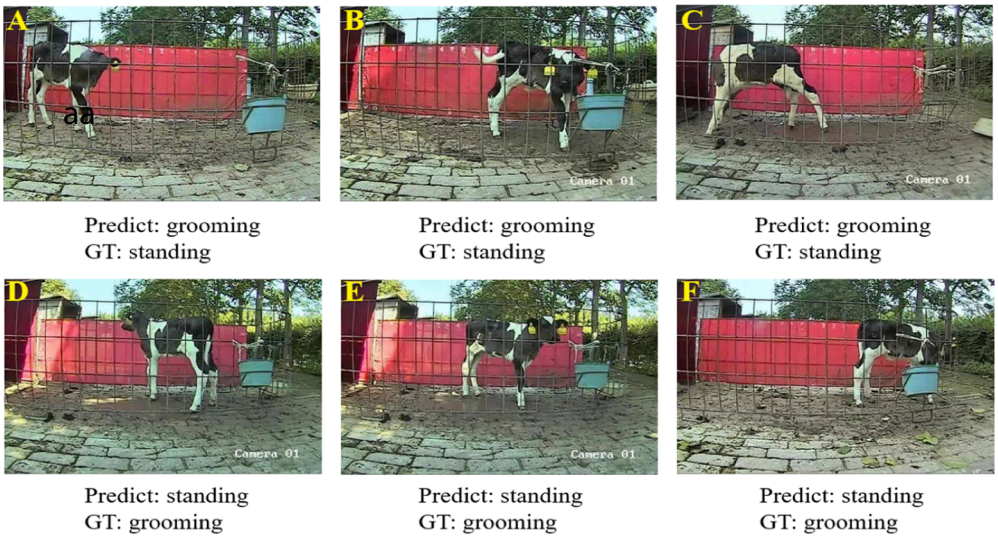



(a) C3D (b) C3D-ConvLSTM

**Fig. 5.** Confusion matrix of behaviour classification in calf dataset.




(a) True classification examples




(b) False classification examples

**Fig. 6.** Behaviour classification results of the proposed C3D-ConvLSTM in calf dataset.

**Table 4**
Adult cow behaviour classification results (%).

| Methods | Accuracy | Precision | Recall | $F_1$ |
|---|---|---|---|---|
| Inception-V3 (frame) | 54.36 | ‡ | ‡ | ‡ |
| SimpleRNN | 70.00 | 76.83 | 57.27 | 65.62 |
| LSTM | 73.33 | 62.03 | 60.71 | 61.36 |
| BilSTM | 75.00 | 75.69 | 68.83 | 72.10 |
| C3D | 83.33 | 83.16 | 68.26 | 74.97 |
| C3D-ConvLSTM | **86.67** | 86.13 | 74.06 | 79.64 |

Note: ‡ means the value is below 50%.

would all affect the accuracy of behaviour classification.

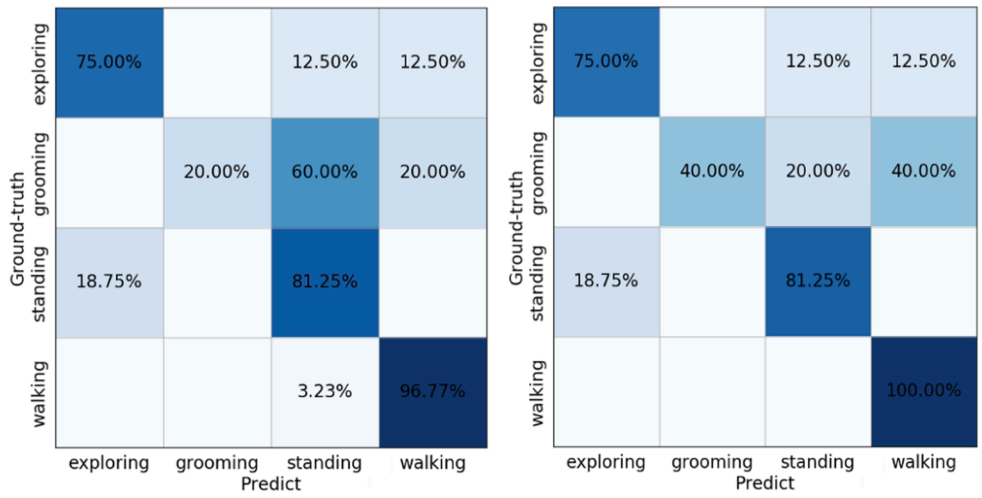
### 5.3. Sequence length for behaviour classification

The impact of different video lengths (i.e. 10, 20, 30 and 40 frame length) with respect to behaviour classification accuracy are provided in Fig. 9. The highest behaviour classification accuracy of calf and adult cow was at 30-frame, the corresponding classification behaviour accuracies were 90.32% and 86.67% respectively, which outperformed 10-frame (76.34% and 76.67%) and 20-frame (81.67% and 78.33%) video length. Additionally, in the stage of 0 to 30 frames, accuracy of both calf and adult cow improved with increased sequence length as more useful spatial–temporal features could be extracted. However, increasing the duration of video past 30 frames appears unhelpful.

Additionally, it can be seen that the overall behaviour classification accuracy for the calf is slightly better than that for the adult cow. The possible reason is that the calf motion is more actively than that of the cow. Therefore, some behaviour patterns such as exploring or walking are more prone to be distinguished from other behaviours.

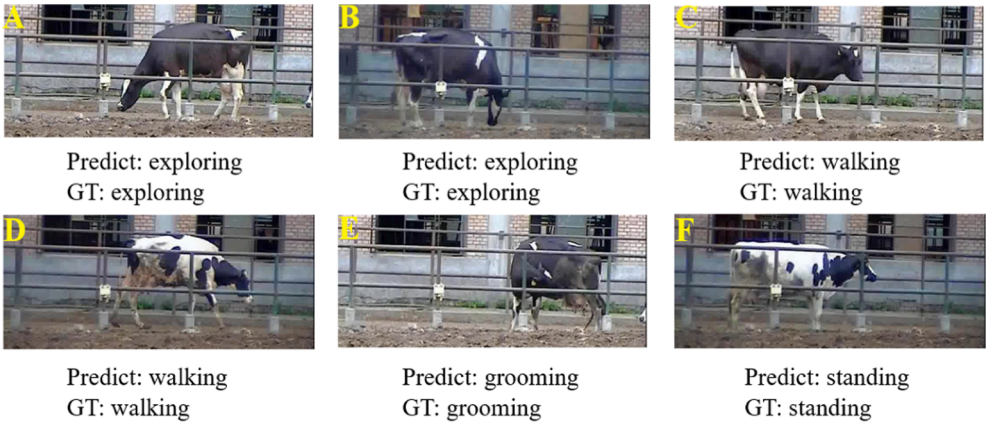
## 6. Discussion

Video-based behaviour recognition commonly used RNN, LSTM and their derived algorithms to extract 2D or 3D visual features (e.g. spatial or location information) (Chen et al., 2021). Compared with image-based behaviour classification methods, video based approach could

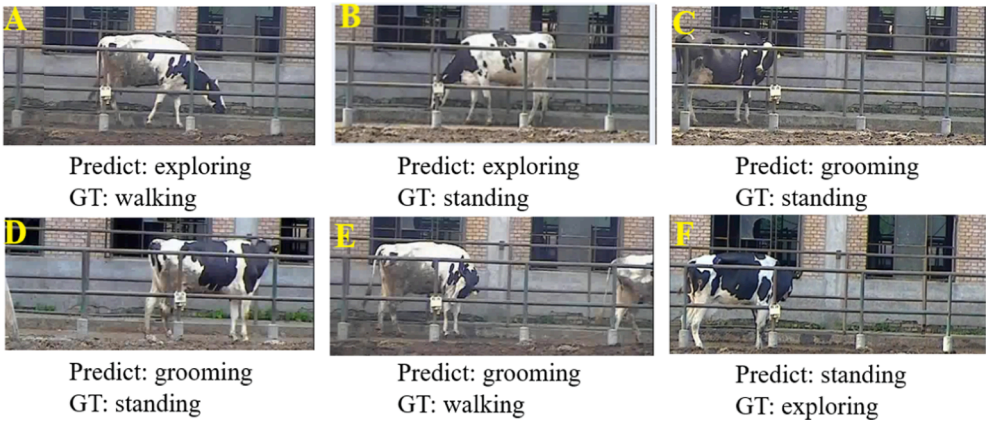



(a) C3D (b) C3D-ConvLSTM

**Fig. 7.** Confusion matrix of behaviour classification on cow dataset.




(a) True classification examples




(b) False classification examples

**Fig. 8.** Behaviour classification results of the proposed C3D-ConvLSTM in cow dataset.

extract temporal characteristics of behaviour, which can better represent animal behaviour, because behaviour is composed of time-series activities (Fuentes et al., 2020). Yin et al. (2020) proposed EfficientNet-LSTM for the recognition of single cow's motion behaviours. Peng et al. (2020) proposed LSTM-RNN based approach to classify dam behaviour patterns. Li et al. (2020a) utilized CNN to assess pullet

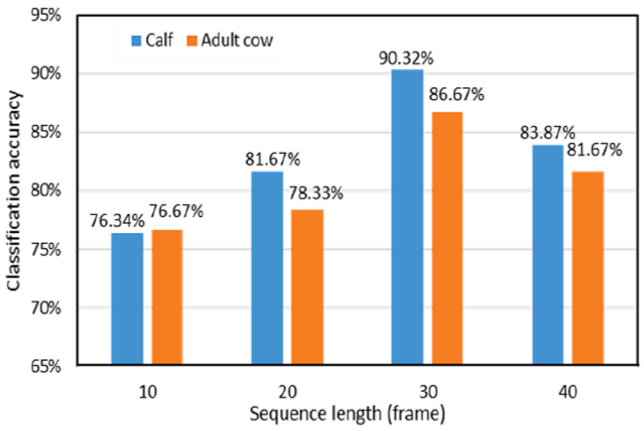



**Fig. 9.** Behaviour classification accuracy w.r.t. sequence length.

drinking behaviours. Guo et al. (2021) extracted 2D features from videos, and utilized BiGRU-attention to capture key spatial–temporal features to improve animal behaviour classification accuracy in precision livestock farming. However, most of these behaviour classification studies focus on the animals with same growth stage, which lacks the robustness.

In this study, two datasets of adult cow and calf were used to explore the applicability of the proposed model. Our proposed C3D-ConvLSTM combines 3D CNN and ConvLSTM to capture the spatial–temporal features, and improves the behaviour classification performance (seen in Table 3 and 4). It can be also noted that C3D-ConvLSTM achieved above 86% behaviour classification accuracy on calf and cow dataset, which is favorable for behaviour classification to different growth stages of dairy cows. The reasons that affect the classification results are: physical discomfort (e.g. lame); environmental factors (e.g. light, obstruction, etc.); the posture or angle of the cow, etc. In addition, the majority of spatio-temporal methods focus on animals performing actions but ignore the relationship between animal and their surrounding objects (Ji et al., 2020; Girdhar et al., 2019).

In terms of practical farming applications, the proposed C3D-ConvLSTM deep learning architecture does not need all input videos have the same length. The long videos will be automatically cropped into same-length clips, while the short video will be expanded by using the last frame with the same padding. For future work, we plan to get depth information using RGB-D camera or 3D LiDAR to further obtain semantic information of behaviors to improve the classification accuracy of transition behaviors or similar behaviors.

## 7. Conclusion

The research of cow behaviour could help improve the precision farm management, cow comfort and well-being. In this work, we proposed a deep learning framework which intelligently combines C3D and ConvLSTM to realize a high-accuracy classification of cow behaviours. The proposed C3D-ConvLSTM based approach achieved an accuracy of 90.32%, a precision of 90.62%, a recall of 88.76%, and an $F_1$ score of 89.68% in calf behaviour classification, and achieved an accuracy of 86.67%, a precision of 86.13%, a recall of 74.06%, and an $F_1$ score of 79.64% in adult cow behaviour classification. These yielded values are higher than those of Inception-V3, SimpleRNN, LSTM, BiLSTM and C3D. The results illustrate that the proposed C3D-ConvLSTM based approach could enhance the classification ability through spatial–temporal feature extraction. In addition, by analyzing the impact of different video lengths on the accuracy of behaviour classification, it is found that the classification accuracy is the best when the video length is 30 frames (compared with 10, 20 and 40 frames). Overall, our proposed C3D-ConvLSTM approach provides a video based deep learning framework for classification of cow behaviours in PLF, and also supply a technical reference for other animal farming (e.g. pig, sheep).

## Declaration of Competing Interest

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

## Acknowledgments

The authors express their gratitude to Kaixuan Zhao for their help with data collection. Also particular thanks to other members of the team for their involvement and efforts in the whole experiment organization and information collection. The authors also acknowledge the support by the project: National Natural Science Foundation of China (Grant No. 61473235) and the National Key Technology R&D Program of China (Grant No. 2017YFD0701603).

## References

Achour, B., Belkadi, M., Aoudjit, R., Laghrouche, M., Lalam, M., Daoui, M., 2021. Classification of dairy cows' behavior by energy-efficient sensor. Journal of Reliable Intelligent. Environments 1–18.

Andriamandroso, A.L.H., Lebeau, F., Beckers, Y., Froidmont, E., Dufrasne, I., Heinesch, B., Dumortier, P., Blanchy, G., Blaise, Y., Bindelle, J., 2017. Development of an open-source algorithm based on inertial measurement units (imu) of a smartphone to detect cattle grass intake and ruminating behaviors. Computers and electronics in agriculture 139, 126–137.

Arablouei, R., Currie, L., Kusy, B., Ingham, A., Greenwood, P.L., Bishop-Hurley, G., 2021. In-situ classification of cattle behavior using accelerometry data. Computers and Electronics in Agriculture 183, 106045.

Balch, T., Dellaert, F., Feldman, A., Guillory, A., Isbell, C.L., Khan, Z., Pratt, S.C., Stein, A. N., Wilde, H., 2006. How multirobot systems research will accelerate our understanding of social animal behavior. Proc. IEEE 94, 1445–1463.

Benaissa, S., Tuyttens, F.A., Plets, D., De Pessemier, T., Trogh, J., Tanghe, E., Martens, L., Vandaele, L., Van Nuffel, A., Joseph, W., et al., 2019. On the use of on-cow accelerometers for the classification of behaviours in dairy barns. Research in veterinary science 125, 425–433.

Chen, C., Zhu, W., Norton, T., 2021. Behaviour recognition of pigs and cattle: Journey from computer vision to deep learning. Computers and Electronics in Agriculture 187, 106255.

Chollet, F., et al., 2018. Keras: The python deep learning library. Astrophysics Source Code Library.

Fuentes, A., Yoon, S., Park, J., Park, D.S., 2020. Deep learning-based hierarchical cattle behavior recognition with spatio-temporal information. Computers and Electronics in Agriculture 177, 105627.

Girdhar, R., Carreira, J., Doersch, C., Zisserman, A., 2019. Video action transformer network. In: In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 244–253.

Gu, J., Wang, Z., Gao, R., Wu, H., 2017a. Cow behavior recognition based on image analysis and activities. International Journal of Agricultural and Biological Engineering 10, 165–174.

Gu, J., Wang, Z., Gao, R., Wu, H., 2017b. Recognition method of cow behavior based on combination of image and activities. Nongye Jixie Xuebao/Transactions of the Chinese Society for Agricultural Machinery 48, 145–151.

Guo, X., Zhang, H., Yang, H., Xu, L., Ye, Z., 2019a. A single attention-based combination of cnn and rnn for relation classification. IEEE Access 7, 12467–12475.

Guo, Y., He, D., Chai, L., 2020. A machine vision-based method for monitoring scene-interactive behaviors of dairy calf. Animals 10, 190.

Guo, Y., Qiao, Y., Sukkarieh, S., Chai, L., He, D., 2021. Bigru-attention based cow behavior classification using video data for precision livestock farming. Transactions of the ASABE 64, 1823–1833.

Guo, Y., Zhang, Z., He, D., Niu, J., Tan, Y., 2019b. Detection of cow mounting behavior using region geometry and optical flow characteristics. Computers and Electronics in Agriculture 163, 104828.

Itakura, K., Saito, Y., Suzuki, T., Kondo, N., Hosoi, F., 2019. Classification of soymilk and tofu with diffuse reflection light using a deep learning technique. AgriEngineering 1, 235–245.

Ji, J., Krishna, R., Fei-Fei, L., Niebles, J.C., 2020. Action genome: Actions as compositions of spatio-temporal scene graphs. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10236–10247.

Karim, F., Majumdar, S., Darabi, H., Harford, S., 2019. Multivariate lstm-fcns for time series classification. Neural Networks 116, 237–245.

Li, G., Hui, X., Chen, Z., Chesser Jr, G.D., Zhao, Y., 2021. Development and evaluation of a method to detect broilers continuously walking around feeder as an indication of restricted feeding behaviors. Computers and Electronics in Agriculture 181, 105982.

Li, G., Ji, B., Li, B., Shi, Z., Zhao, Y., Dou, Y., Brocato, J., 2020a. Assessment of layer pullet drinking behaviors under selectable light colors using convolutional neural network. Computers and Electronics in Agriculture 172, 105333.

Li, J., Wu, P., Kang, F., Zhang, L., Xuan, C., 2018a. Study on the detection of dairy cows' self-protective behaviors based on vision analysis. Advances in Multimedia, 2018.
Li, W., Qi, F., Tang, M., Yu, Z., 2020b. Bidirectional lstm with self-attention mechanism and multi-channel features for sentiment classification. Neurocomputing.
Li, X., Stevens, A., Greenberg, J.A., Gehm, M.E., 2018b. Single-shot memory-effect video. Scientific reports 8, 1–8.
Liu, D., He, D., Norton, T., 2020. Automatic estimation of dairy cattle body condition score from depth image using ensemble model. Biosyst. Eng. 194, 16–27.
Meunier, B., Pradel, P., Sloth, K.H., Cirié, C., Delval, E., Mialon, M.M., Veissier, I., 2018. Image analysis to refine measurements of dairy cow behaviour from a real-time location system. Biosystems engineering 173, 32–44.
Moran, J., Doyle, R., 2015. Cow talk: Understanding dairy cow behaviour to improve their welfare on Asian farms. CSIRO PUBLISHING.
Ordó nez, F., Roggen, D., 2016. Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition. Sensors 16, 115.
Ouyang, X., Xu, S., Zhang, C., Zhou, P., Yang, Y., Liu, G., Li, X., 2019. A 3d-cnn and lstm based multi-task learning architecture for action recognition. IEEE Access 7, 40757–40770.
Pavlovic, D., Davison, C., Hamilton, A., Marko, O., Atkinson, R., Michie, C., Crnojević, V., Andonovic, I., Bellekens, X., Tachtatzis, C., 2021. Classification of cattle behaviours using neck-mounted accelerometer-equipped collars and convolutional neural networks. Sensors 21, 4050.
Peng, Y., Kondo, N., Fujiura, T., Suzuki, T., Ouma, S., Yoshioka, H., Itoyama, E., et al., 2020. Dam behavior patterns in japanese black beef cattle prior to calving: Automated detection using lstm-rnn. Computers and Electronics in Agriculture 169, 105178.
Peng, Y., Kondo, N., Fujiura, T., Suzuki, T., Yoshioka, H., Itoyama, E., et al., 2019. Classification of multiple cattle behavior patterns using a recurrent neural network with long short-term memory and inertial measurement units. Computers and Electronics in Agriculture 157, 247–253.
Qiao, Y., Kong, H., Clark, C., Lomax, S., Su, D., Eiffert, S., Sukkarieh, S., 2021. Intelligent perception-based cattle lameness detection and behaviour recognition: A review. Animals 11, 3033.
Qiao, Y., Kong, H., Clark, C., Lomax, S., Su, D., Eiffert, S., Sukkarieh, S., 2021. Intelligent perception for cattle monitoring: A review for cattle identification, body condition score evaluation, and weight estimation. Computers and Electronics in Agriculture 185, 106143.
Qiao, Y., Truman, M., Sukkarieh, S., 2019. Cattle segmentation and contour extraction based on mask r-cnn for precision livestock farming. Computers and Electronics in Agriculture 165, 104958.
Rahman, A., Smith, D., Little, B., Ingham, A., Greenwood, P., Bishop-Hurley, G., 2018. Cattle behaviour classification from collar, halter, and ear tag sensors. Information processing in agriculture 5, 124–133.
Riaboff, L., Poggi, S., Madouasse, A., Couvreur, S., Aubin, S., Bédère, N., Goumand, E., Chauvin, A., Plantier, G., 2020. Development of a methodological framework for a robust prediction of the main behaviours of dairy cows using a combination of machine learning algorithms on accelerometer data. Comput. Electron. Agric. 169, 105179.
Ronghua, G., JingQiu, G., Jubao, L., 2017. Cow behavioral recognition using dynamic analysis. In: Smart Grid and Electrical Automation (ICSGEA), 2017 International Conference on. IEEE, pp. 335–338.
Salau, J., Krieter, J., 2020. Analysing the space-usage-pattern of a cow herd using video surveillance and automated motion detection. Biosyst. Eng. 197, 122–134.
Smith, D., Little, B., Greenwood, P.I., Valencia, P., Rahman, A., Ingham, A., Bishop-Hurley, G., Shahriar, M.S., Hellicar, A., 2015. A study of sensor derived features in cattle behaviour classification models. In: 2015 IEEE SENSORS. IEEE, pp. 1–4.
Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z., 2016. Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2818–2826.
Tamura, T., Okubo, Y., Deguchi, Y., Koshikawa, S., Takahashi, M., Chida, Y., Okada, K., 2019. Dairy cattle behavior classifications based on decision tree learning using 3-axis neck-mounted accelerometers. Animal Sci. J. 90, 589–596.
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, pp. 4489–4497.
Weizheng, S., Fei, C., Yu, Z., Xiaoli, W., Qiang, F., Yonggen, Z., 2019. Automatic recognition of ingestive-related behaviors of dairy cows based on triaxial acceleration. Information Processing in Agriculture.
Wu, X., Wu, Z., Zhang, J., Ju, L., Wang, S., 2020. Salsac: A video saliency prediction model with shuffled attentions and correlation-based convlstm. In: AAAI, pp. 12410–12417.
Xue, T., Qiao, Y., Kong, H., Su, D., Pan, S., Rafique, K., Sukkarieh, S., 2021. One-shot learning-based animal video segmentation. IEEE Trans. Industr. Inf.
Yin, X., Wu, D., Shang, Y., Jiang, B., Song, H., 2020. Using an efficientnet-lstm for the recognition of single cow's motion behaviours in a complicated environment. Comput. Electron. Agric. 177, 105707.