AGRICULTURAL AND FOOD SCIENCE

Agricultural and Food Science (2022) 31: 89–103
https://doi.org/10.23986/afsci.111665

Deep learning image recognition of cow behavior and an open data set acquired near an automatic milking robot

Olli Koskela[1], Leonardo Santiago Benitez Pereira[1,2], Ilpo Pölönen[3], Ilmo Aronen[4] and Iivari Kunttu[1]

[1]HAMK Smart research unit, Häme University of Applied Sciences, Hämeenlinna, Finland
[2]Federal Institute of Santa Catarina, Florianópolis, Brazil
[3]HAMK Bio research unit, Häme University of Applied Sciences, Hämeenlinna, Finland
[4]Lantmännen Feed Oy, Turku, Finland
e-mail: olli.koskela@hamk.fi

Production animals enjoying good health and well-being are more productive and have a higher output quality. Several technical solutions have been used to monitor the animals' welfare: those based on computer vision provide cost-efficient and scalable options. In this work, we performed a continuous two-month image acquisition of cows in front of an automatic milking station and divided the data into ten different classes related to the most important activities appearing in the images. The data consisted of almost 19 hours of video, equivalent to more than 1.7 million still images. Based on these imaged, we then implemented a convolutional neural network classifier to recognize the cows' behavior. The network was tested using cross-validation methodology and achieved an 86% precision rate and 85% recall rate in the classification. The data and the Python program code used in this study are made available. An image data set that directly resembles the harsh conditions inside a barn and can be used for deep learning purposes has not been previously made available.

Key words: animal well-being, video recording

Introduction

Farm animals enjoying good health and well-being are more productive with higher quality yield (Wegner et al. 1976, Berckmans 2014), and the importance of understanding these phenomena has grown as the farm sizes have grown. It is no longer possible for a farmer to follow individual animals; instead, decision making is based more on obtaining averages from certain indicators for the whole herd. Nonetheless, care is shown for each animal individually and measurements are made for individual animals to aggregate herd averages.

Several technical solutions have been used to monitor animal welfare, including the use of digital video cameras (Porto et al. 2015, Ardo et al. 2017, Ter-Sarkisov et al. 2017), depth sensor cameras (Nasirahmadi et al. 2017), sound (Schirmann et al. 2009); three-dimensional accelerometers (Müller et al. 2003, Steeneveld and Hogeveen 2015, Gardenier et al. 2018, Shen et al. 2020); and infrared thermography (de Sousa et al. 2018, Cuthbertson et al. 2019, Xudong et al. 2020, Anagnostopoulos et al. 2021). Computer vision approaches employing video cameras are scalable and low-cost solutions (Banhazi and Tscharke 2016) and have been successfully used to monitor physiological and behavioral parameters related to pre-slaughter stress (Jorquera-Chavez et al. 2019); to detect hoof disease (Gu et al. 2017); to track gait and identify lameness (Gardenier et al. 2018, Jiang et al. 2019a, Kang et al. 2022), to analyze health problems through calculating body condition scores (Zin et al. 2018b, Huang et al. 2019), body structure (Jiang et al. 2019 b, Liu et al. 2020) and faecal monitoring (Atkinson et al. 2020); and to detect aggressive behaviors (Chen et al. 2019). One often used computer vision method is to segment instances to recognize and track individual cows (Guzhva et al. 2018, Ter-Sarkisov et al. 2018, Zin et al. 2018a, Qiao et al. 2019, Shao et al. 2019, Li et al. 2021).

In computer vision applications, deep learning approaches using Convolutional Neural Networks (CNN) are able to achieve state-of-the-art results using only the image data, thus requiring little domain knowledge. Previously trained networks can be transferred from one context to another to increase the information learned by the network (Chollet 2017). Furthermore, after the initial training phase, the CNNs are computationally light to deploy or embed within applications.

Using commercially available security camera, we performed continuous, two-month image acquisition of cows in front of an Automatic Milking Station (AMS) inside a teaching barn in Mustiala, Tammela, Finland. The chosen area serves a specific purpose, milking, and hence, cows are motivated to visit the spot several times per day. From the acquired images, a set of 1 700 000 were classified into 10 scenes that were based on cow-actions available in the

Received 11 October 2021 / Accepted 9 May 2022
The Scientific Agricultural Society of Finland

set and following the guidelines defined in (Boissy et al. 2007, OIE 2019). The Terrestrial Code (OIE 2019), derived by the World Organisation for Animal Health (OIE), is the main international standard for animal welfare and defines several measurables to monitor the impact of design and management in animal welfare. One of the main measurables with respect to dairy cattle is their behavior, including aggressions between animals, decreased feed intake, and altered locomotor behavior and posture. Another measurable is related to the human-animal relationship, such as aggressiveness during handling or keeping excessive distance. Moreover, Boissy et al. (2007) point out the importance of assessing positive emotions: exploration, grooming, and friendly interactions among animals.

To demonstrate a monitoring system for behavioral actions, our aim was to detect cow behavior in a constrained area using a VGG-based CNN using the acquired image data classified into suitable action categories. Our results show the potential of this approach along with several features that need to be considered in applications. The developed CNN developed was tested using Cross-Validation methodology (Faceli et al. 2011), and it achieved an 86% precision rate and an 85% recall rate in the classification task. We have shared the data set used in this article – equivalent to almost 19 hours of video – along with the Python program code to implement the full data software pipeline and the CNNin Benitez Pereira et al. (2020), licensed with the Creative Commons 4.0 Attribute license.

Materials and methods

Data acquisition

For the video data acquisition, we used a commercial security dome camera (VIO-D30, Blaupunkt GmbH, Ingress Protection class IP67). It provides a 100° wide-angle view. The camera captures video when movement is detected and, on average, the video files are about four minutes in duration and include 15 videos per hour. The native video format is an H.264-encoded, 25 frames per second, 1920-by-1080 DAV-format file.

We initially installed the camera on 22 September 2019 and placed it 3.3 m above ground level and 3.3 m left of the front left corner of the AMS housing seen in Figure 1a. To increase variety in the training data set, we changed the camera position 1.0 m further left of AMS and rotated it 3.4° horizontally towards AMS and 1.7° vertically upwards on October 25, 2019. Figure 1b shows an overview of the setting and the region of interest of this study.

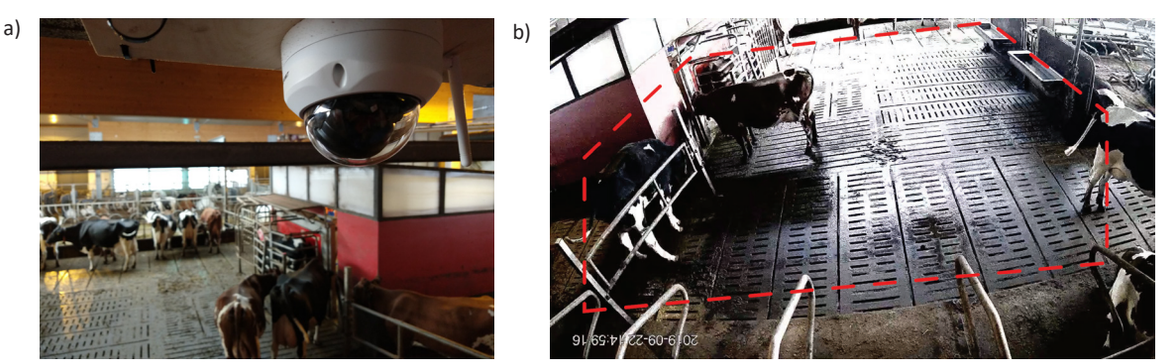

Fig. 1. Overview of the camera installation (a) and (b) region of interest perimeter marked with the red-dotted line. AMS was located in the red housing on the left of region of interest and the region of interest includes the part of AMS accessible to cows. Note that the original image captured by the camera was rotated 180° and, hence, the date label is upside down in the images.

Videos chosen for labeling

Two periods of video footage were labeled, the first for the initial camera position from 22 September to 18 October 2019, and the second for altered camera position from 27 October to 20 November 2019. Starting from the beginning of the period, ten videos from every second day were labeled for both periods. For the first period, the labeled videos were the last video of the hours 03, 06, 08, 10, 11, 13, 14, 15, 17, and 19, and for the second period they were the last video of the hours 03, 08, 09, 10, 12, 13, 14, 15, 17, and 19. Exact staring times of the labeled videos are provided in the supplementary material. The videos were chosen with no a priori information on the days (nor are any major events known afterwards), and they consist of one video from around the sunrise and sunset, multiple videos rather equispaced throughout the day, and one video from the middle of the night,

where the camera had a poor signal (hence only one night video per day). With this design, we intended to cover a reasonable amount of time while still keeping the data set and labeling process feasible. In total, 280 videos were labeled with the average duration of four minutes and nine seconds.

For the labeling and analysis of the video, the native DAV format was converted to MP4 using the FFmpeg library. Some DAV to MP4 conversions failed for no apparent reason. We experienced eight such conversion failures, tracked them, and substituted video from the same hour, but from the next day. Substituted videos are listed in the supplementary material.

Classes used for labeling

The classes consist of one for humans in the image, three for cow-to-cow interactions, two for the number of queuing cows, two for cow activities, and two for no cows in the image. Each frame of the videos was treated as a single image and assigned to one of those classes. Abnormal or undefined situations were assigned to an undefined class (number -1). The defined classes are not mutually exclusive, so if the same scene included two classes (example: one cow is drinking water while a frontal interaction is happening), we adopted a priority scale: just the situation with a higher priority (lower number class) was saved. An overview of the classes in their order of priority is given in Table 1.

Table 1. The classes used in labeling in their order of priority decreasing from top to bottom

code	name	brief description
-1	undefined	Undefined situation or doubt about the correct class due to too much noise
0	human	There are human(s) in the image.
1	interaction frontal	Head on head or neck
2	interaction lateral	Head or body touching the other cow's head or body
3	interaction vertical	Head or other body part over body or head
4	crowded	More than five cows in the interest area
5	drinking	A cow is drinking water.
6	curiosity	Putting head inside the AMS housing.
7	queue	At least one cow is waiting outside the milking station.
8	low visibility	Night or twilight
9	normal	Nothing, laying down, peeing, etc.

The undefined and human classes were prioritized the most because of their importance for further analysis: undefined images were naturally omitted from the CNN training; correct human class prediction is important when considering that the behavior of the cows is different when humans are present. The undefined class was used when the image was very noisy or otherwise there was doubt about the correct class. The human class was assigned when at least 30% of a human body was visible in the image based on the subjective estimate of the labeler.

Interactions between cows were divided into three categories: frontal, lateral, and vertical interactions. In frontal interactions, two cows were interacting and facing each other, as shown schematically in Figure 2a. Frontal interaction does not have to be mutual, only that both of the cows were considered to be are able to react to the approaching interaction before it happens. In lateral interaction, a cow is pushing, kicking, or heading into another cow, whether reciprocally or not. Four schematic situations are shown in Figure 2b. In vertical interaction, a cow has positioned any of its body parts over another cow. In practice, the body parts that can be positioned over another cow are the head and torso, thus vertical interaction includes mounting behavior. The schematics of vertical interactions are shown in Figure 2c.

A crowded situation was assigned when there were more than five cows entirely inside the area next to the AMS. That area is delimited by considering the scope of the data set (cow behavior near AMS) and is defined as the area inside the dashed perimeter in Figure 1b.

Drinking was only assigned if the cow was actively engaged in drinking. It is very common that the cow just stays near the water drinker, or just licks the metallic part of it, and these situations were not included in this class.

Curiosity was assigned when a cow had its head at least partially in the opening near the entrance of the AMS. This action was considered a natural exploring of the environment (a relevant indicator of positive affective states according to Boissy et al (2007), especially since the milking reward feed is kept within the housing. A queue represented a situation in which at least one cow was waiting outside the milking station. The queue could vary in size and organization, with the only constraint being that the cows needed to be facing the milking station. That is, it involved a scene where some cow was near the milking station; however, cow apparently just walking by the station was not included under this class. This class is subjective and dependent on the personal interpretation of the person doing the annotation, so a more precise task would require additional labeling and/or redefinition of the classes.

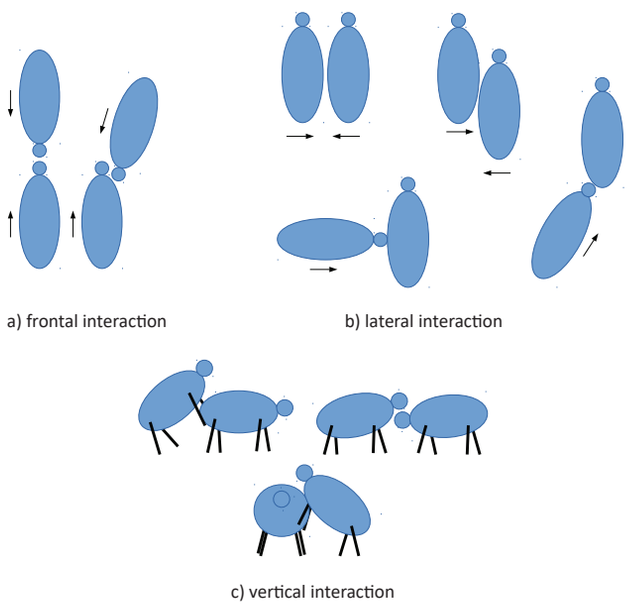

Fig. 2. Schematic examples of interaction classes: (a) frontal, (b) lateral, and (c) vertical

Low visibility and normal are both situations defined mainly by the absence of the actions described above. Low visibility is used during the night, when the lights of the barn are dimmed and the video footage is almost completely blank. These images are labeled in this class. For most of applications, these images can probably just be discarded, but we decided to keep them for completeness. A normal situation is the complement of all the other classes and could be interpreted as a binary classification with cases in which something was happening versus when nothing was happening.

Labeling process

The videos were annotated so that each frame was considered a single image and assigned to exactly one class. We developed Python software that takes the MP4 video, plays the video (with options to rewind, advance, and pause the video), and saves a label correspondent to each frame of the video. Label input is given by the user pressing a corresponding key. Idea was based on a tool developed by (van den Berg et al. 2011).

The assigned classes were written as text files in JSON (JavaScript Object Notation) format. We will refer to the JSON class files as pointer tables. Each line of the pointer table contains a class label for one frame in a triplet: (video file name, frame number in the video, assigned class). For example, (1569103140, 1, 8) would be the first frame of video 1569103140.mp4 assigned to class 8, low visibility. To ensure the reproducibility of the results, we stored a pre-shuffled pointer table, which was used in our experiments.

A thousand images required approximately one minute to label, including all the tasks involved in the annotation process. The videos occupy 40.7 GB stored in mp4 format, while the labels occupy 82.9 MB stored in plain text file.

CNN architecture and training

To accomplish the classification task, we used a CNN that was implemented using Python and Keras. The choice of network was based on a heuristic altering the VGG (Simonyan and Zisserman 2014) architecture with performance evaluated using the cross-validation methodology. The VGG architecture was originally developed to perform image classification and object localization among 1000 classes. Compared to other previously published architectures using large receptive fields (size of convolution filter), for instance 11-by-11 with stride 4 in (Krizhevsky et al. 2012), the VGG used very small (3-by-3) receptive fields with stride 1 and a higher number of layers, that is to say, it is a deeper neural network. A filter size of 3-by-3 is the smallest size at which the notion of left/right, up/down, and center can be captured, making the decision function more discriminative. Other often used networks in precision farming include ResNet, DenseNet, Inception V3, and MobileNet, but no significant difference beforehand has been demonstrated (Mahmud et al. 2021).

We found the best performance with the architecture shown in Figure 3. It consists of seven convolutional layers with the ReLu activation function (Chollet 2017) and different numbers of filters. Each convolutional layer is followed by a pooling layer to reduce the dimension. The final classifier is implemented with two fully connected layers with the ReLu activation function, each one regularized with a 20% dropout rate (random drop of neurons from the neural network during training, helping to prevent overfitting (Srivastava et al. 2014). The last layer is a fully connected layer with a softmax activation function (Chollet 2017), used to produce a probability distribution as output, and it has ten output neurons constituting the probabilities that an image belongs to any of the classes. In total, the architecture has 20 layers and the Python implementation is provided alongside the data set in the supplementary material.

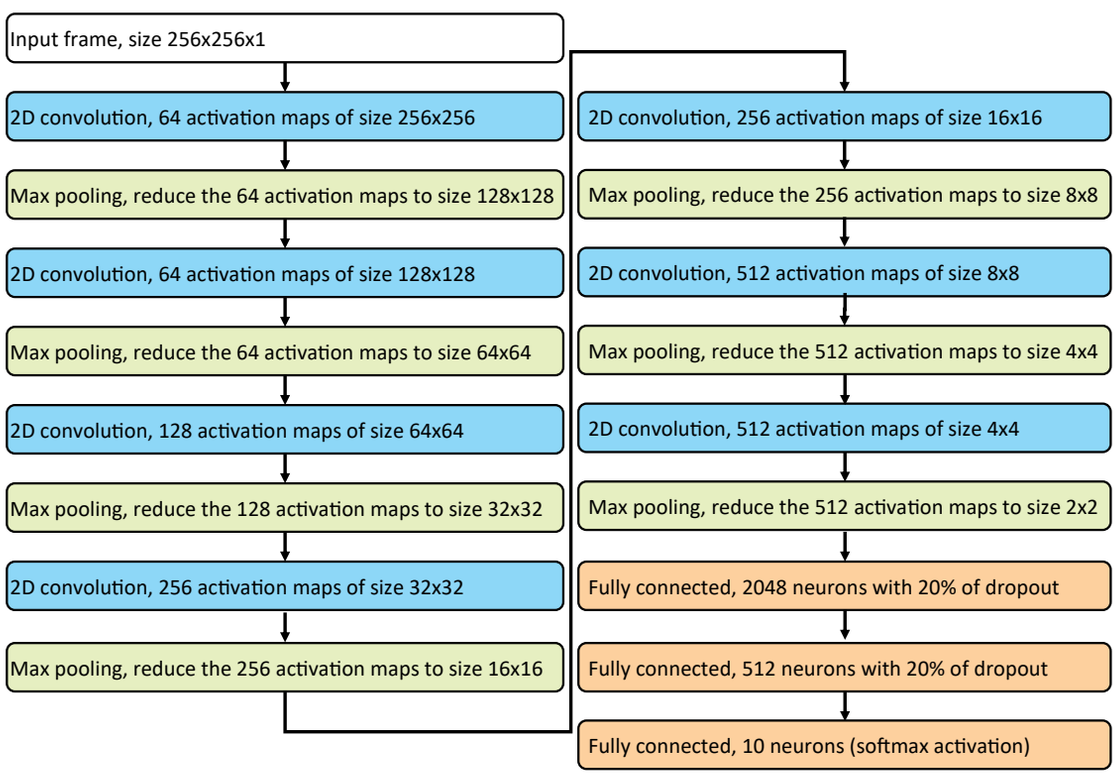

Fig. 3. Architecture of the CNN used in this article, consisting of seven convolutional and max-pooling layer pairs and three pooling layers

The loss function chosen for our purposes was categorical cross-entropy loss because of the multiclass classification nature of the problem. The system was optimized with the Adam algorithm and evaluated with the accuracy metric.

We trained two CNNs: one with a subset of 100 000 images randomly sampled directly from the labeled data and one with 49 900 sampled such that each class is represented with an equal number of randomly included images. The lowest number of images occurred in the class vertical interaction, which thus defined the number 4990 images for each of the ten classes. These CNNs are later referred to as unbalanced and balanced experiments, respectively. The exact images are listed in the data set (Benitez Pereira et al. 2020).

The CNNs were trained with images average down-scaled to a 256-by-256 resolution and converted to grayscale by averaging the intensities of the three channels. The full data set was shuffled once and recorded in shuffled tables, which are provided in the supplementary material for reproducibility. In the unbalanced experiment, we used 80 000 (80%) images for training, 10 000 (10%) for validation, and 10 000 (10%) for testing. In the balanced experiment, the number of images were 24 950 (50%) for training, 12 475 (25%) for validation, and 12 475 (25%) for testing. The network was trained for 25 epochs in unbalanced experiment and 27 in balanced experiment.

Before using the images to train the network, we augmented them by applying random geometrical transformations using the library Albumentations (Buslaev et al. 2020). This made it possible to have more variation in the data (thus, yielding with less bias towards over-fitting); for instance, all the data was collected during fall, which is not enough to represent the changes in daylight throughout the year in Finland, and therefore it is beneficial to augment the data by varying the exposure. The applied augmentations are summarized in Table 2, which also shows the range of the augmentation adjust parameter. The value of the adjust parameter is chosen randomly from a uniform distribution within the given range, and it defines the intensity of the augmentation. Each type of augmentation has 50% change of appearance.

Table 2. Training set augmentation transformations and ranges

transformation	range
contrast	0% – 30%
gamma	0% – 30%
brightness	0% – 50%
shift	0% – 5%
scale	0% – 5%
rotate	0° – 10°

We chose ranges of the adjust parameter that would be within reasonable limits considering the environment. For example, excessive rotation and shifting were avoided. Also, shifting, scaling, and rotation preserve the boundary of the interest area (defined in Figure 1 (b)) to be included in the augmented image. Contrast, gamma, and brightness adjust parameter ranges were chosen so that even the most augmented image might represent a possible image from the camera. The chosen range of the rotation parameter was a little higher than the modification done in the middle of the experiment (3.4°). In case of semantic segmentation, we suggest the reader to see Qiao et al. (2020) on advanced augmentation of contour label which were not used in this work though.

To validate the trained network, we used cross-validation methodology (Russell and Norvig 2010). Cross-validation methodology divides the data into three independent data sets, from which one is used for training, one to validate the training, and one to test the eventual result. The prediction results of the network from the test set were numerically evaluated using precision, recall and F1-score metrics separately for each class. Precision and recall are defined as

$$\text{precision} = \frac{\#\text{true positives}}{\#\text{true positives} + \#\text{false positives}}$$

and

$$\text{recall} = \frac{\#\text{true positives}}{\#\text{true positives} + \#\text{false negatives}}$$

where true positives are the correct classifications of the evaluated class, false positives the false classifications to the evaluated class, and false negatives the wrong classifications of the evaluated class to another class. F1-score is computed for each class based on their precision and recall as

$$\text{F1-score} = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}.$$

One additional evaluation metric is the Top-3 accuracy, which considers the prediction right when one of the three most confidently predicted classes is equal to the true label, instead of only the most confident one.

Software and hardware used

All computations were performed at an HP Z4 G4 workstation with a NVIDIA® Quadro® P2000 (5 GB GDDR5 dedicated) graphics processing unit, 16 GB DDR4-2400 random access memory, and an Intel® Core™ i7-7820X processor. The software used included Python (version 3.6.9), Keras (version 2.2.4) with backend TensorFlow (version 1.14.0), Keras-vis toolkit (version 0.4.1), Albumentations image augmentation library (version 0.4.2) (Buslaev et al. 2020), and opencv-python (version 3.4.2.17) computer vision library (Bradski 2000).

Ethical statement

During the research, the cows were cared for according to the effective laws and regulations in Finland. We did not intervene with animals. Persons were informed with signs upon entering the acquisition area about the research data acquisition project. The protocol for disconnecting the acquisition of images while visiting the area was given in the signs. Any data including persons have not been published.

Results

Labeled data set

The data set consisted of 280 videos, with a mean duration of 249 seconds (4 minutes and 9 seconds) and a standard deviation of 78 seconds, equivalent to 1 700 660 valid frames (labeled with a class other than -1). Most of the labels, 40.7%, belong to the queue class, while just 0.3% belong to the interaction vertical class, as shown in Figure 4. The relatively large percentage for the queue class is natural given that the monitored region was primarily constructed for entering the AMS. Figure 5 shows one sample frame for each class labeled in this data set.

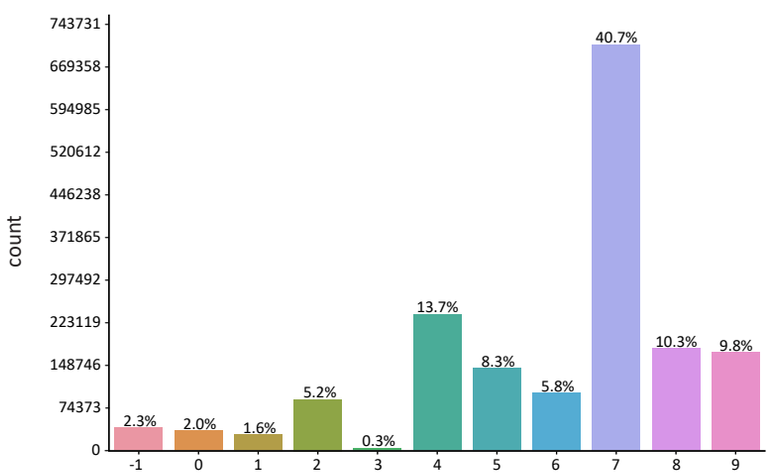

Fig. 4. Distribution of classes throughout the entire data set

The mean number of classes per video was 3.78, with a standard deviation of 1.68. Of all the videos, 61.1% contain at least one interaction event (classes 1, 2, or 3), and 9.6% contain at least one human image. Figure 6 illustrates the temporal distribution of the classes on a video (30 September 2019, video from hour 08), and Figure 7 illustrates daily distribution (12 October 2019).

We conducted a small test to see if the classification of images agreed between two persons. The test was done once for data on a single day, and it resulted in 87.7% agreement in the labeled classes. To avoid any discrepancies, all data in the supplementary, published set was labeled by the same person.

Fig. 5. Sample image of each class

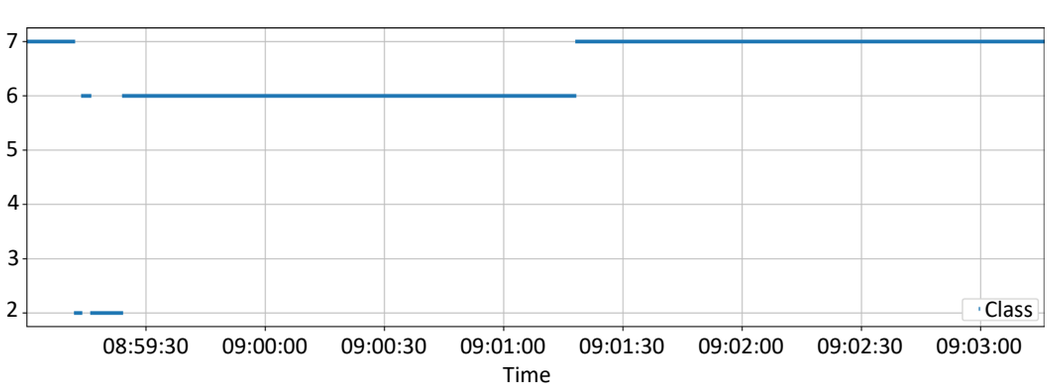

Fig. 6. Temporal distribution of the classes in a sample video (30 September 2019, video from hour 08)

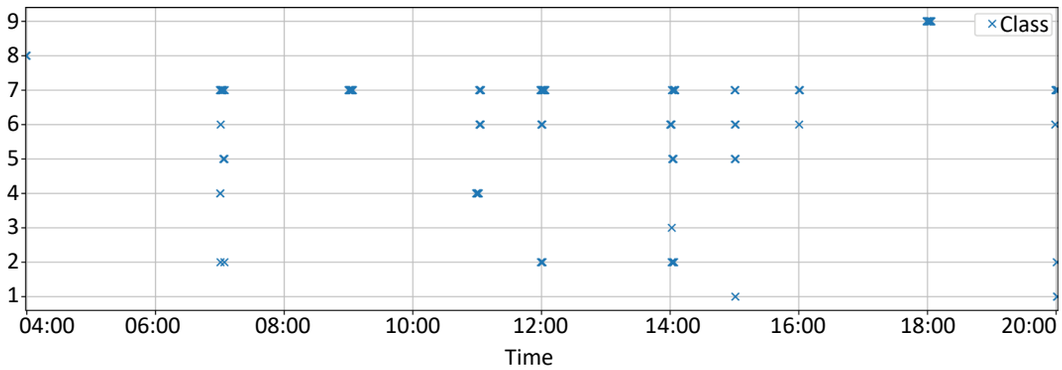

Fig. 7. Temporal distribution of the classes on a typical day (12 October 2019)

Classification experiments with CNNs

Based on the cross-validation evaluation, the CNN in unbalanced experiment achieved an 86% precision rate and an 85% recall rate (mean over all classes, weighted by the number of examples in each class) and the CNN in balanced experiment achieved a 78% precision rate and 79% recall rate. Table 3 shows the precision, recall, F1-score, and support (number of samples in this class) for each class, as well as their weighted average (mean over all classes, weighted by the number of examples in each class) and the overall accuracy. Figure 8 illustrates these results in the form of a confusion matrix, while Figure 9 shows the last eight epochs of training, where the final precision rate was reached and further training shows no improvement. The classes that where most benefited by the balanced experiment were the ones with little representation in the unbalanced experiment (the interaction classes and the human class), but a considerable amount of errors still happen between the classes interaction frontal and interaction vertical, presumably because the two situations are visually similar.

Table 3. Model evaluations

	unbalanced experiment				balanced experiment			
class	precision	recall	F1	support	precision	recall	F1	support
human	91%	50%	65%	224	84%	90%	87%	1252
interaction frontal	100%	27%	43%	147	83%	75%	78%	1290
interaction lateral	79%	28%	42%	510	66%	57%	61%	1264
interaction vertical	100%	9%	16%	23	90%	98%	94%	1267
crowded	86%	96%	91%	1455	68%	83%	75%	1210
drink	82%	74%	78%	869	74%	82%	78%	1237
curiosity	99%	31%	47%	5817	70%	73%	71%	1298
queue	81%	97%	88%	4160	55%	37%	44%	1120
low visibility	100%	100%	100%	1005	99%	100%	99%	1195
normal	92%	95%	94%	1020	89%	91%	90%	1231
accuracy			85%	10 000			79%	12 464
weighted average	86%	85%	83%	10 000	78%	79%	78%	12 464

The different forms of evaluating the model allow for different interpretations (Faceli et al. 2011) and comparison with other works. For our application, the F1-score weighted average is the most important metric because it balances precision and recall while taking into consideration the uneven distribution of the classes in the unbalanced experiment.

While Top-3 metric is not common, it makes sense in our application, where one situation can be very subtle but still of interest. With this metric, the CNN of unbalanced experiment has a Top-3 accuracy of 97% and the CNN in balanced experiment 95%.

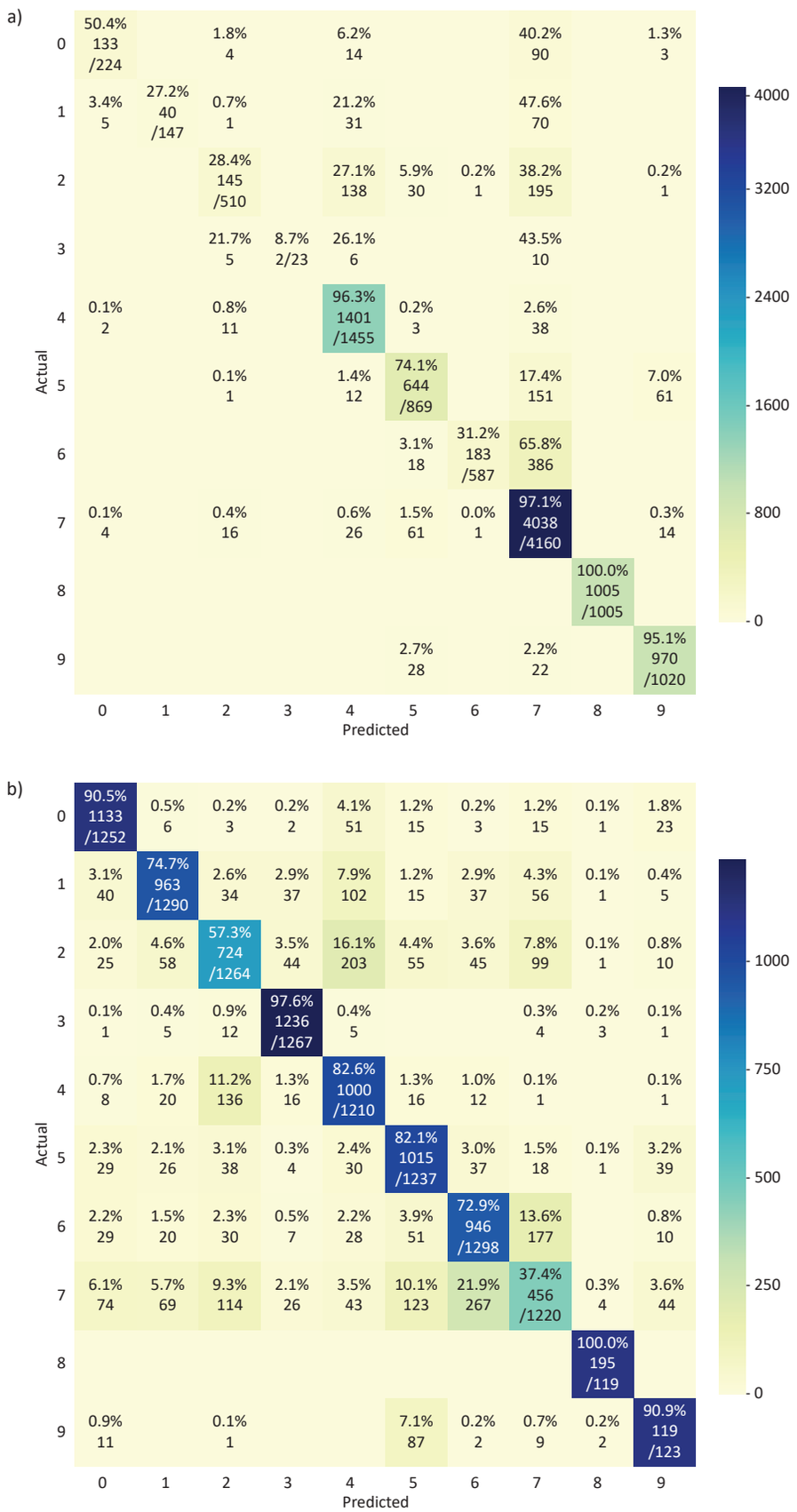

Fig. 8. Confusion matrices for the (a) unbalanced and (b) balanced experiments

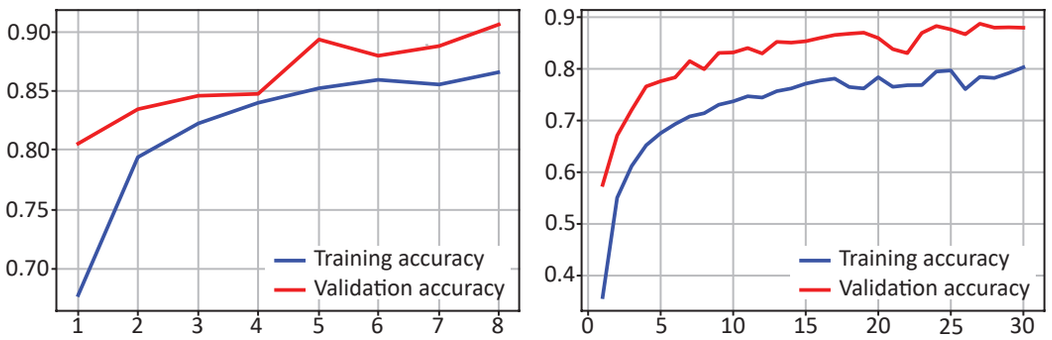

Fig. 9. Accuracy of the training and validation (a) during the last eight epochs of training out of a total of 25 epochs in the unbalanced experiment and (b) all 30 epochs in the balanced experiment

CNN's heatmap activation

A wide array of techniques exists for visualizing and interpreting CNNs (Chollet 2017). As pointed out by (Selvaraju et al. 2019) these visualizations are important because "in order to build trust in intelligent systems and move towards their meaningful integration into our everyday lives, it is clear that we must build 'transparent' models that have the ability to explain why they predict what they predict".

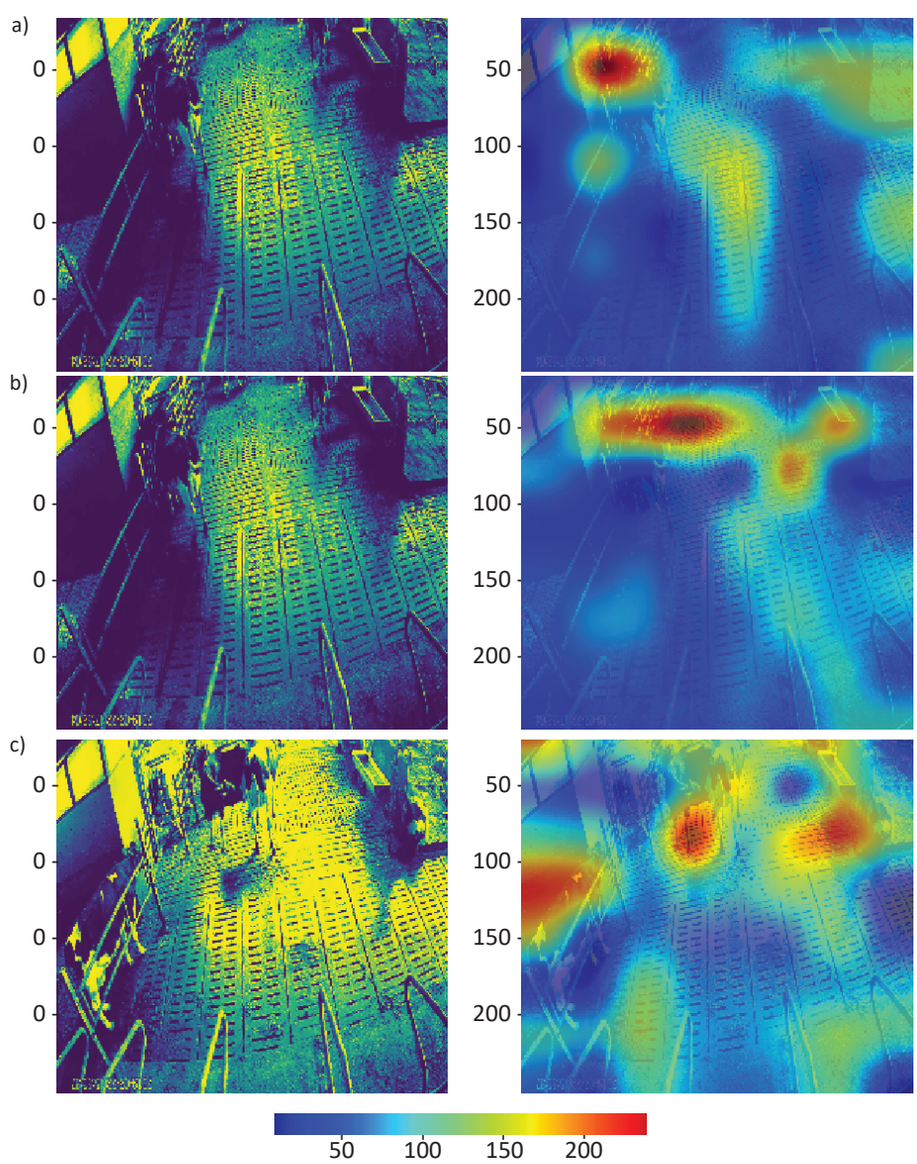

Fig. 10. Images (a) and (b) are from video 1569153540.mp4 frame 6229, classes curiosity and drink, respectively. Image (c) is from video 1569146340 frame 489, class interaction lateral. Note that the alignment between the heatmap and image is not perfect.

In this study, we present heatmaps of class activation based on the Grad-CAM technique (Selvaraju et al. 2019), which shows what parts of the image are important to a given class. To generate meaningful visualizations, we applied the Grad-CAM technique to videos different than the ones the network was trained with, and the penultimate convolutional layer (conv2d_6) was used to generate the visualizations. We used the Keras-vis tool kit, and as recommended in its documentation, we changed the last layer's activation function from softmax to linear. Three selected results obtained from the unbalanced experiment are shown in Figure 10 where it is possible to see that the network correctly identifies the position of the objects of interest for the given class: the curious exploration happening in the top left corner (Fig. 10a), the cow drinking water in the top right corner (Fig. 10b), and two cows interacting in the center of the image (Fig. 10c). However, it also has strong activation in other regions, for instance the false recognition near the cow that is not drinking water in Figure 10b.

Discussion

In this study, we have presented an approach aiming to enable long-term analysis of the behavior of cows in real time in their living environment. We deployed a commercially available security camera that was IP-classified and suitable for the harsh environment of the barn. Also, data transferring was implemented with readily available cloud services.

Our results demonstrate the potential of deep learning in classifying the actions of cows. However, during the project we also identified many pitfalls of this robust, but rather simplistic, approach. No prior works using such a wide number of scenes existed, and hence, we chose to use supervised learning to have better a understanding of the results. In our experimental setup, we decided to limit our study to only a small portion of the barn usually having 50 to 60 cows engaged in milking. This is the only section of the barn where AMS-related events could occur, which was the motivation behind the collection of the data set.

The labeling process was found efficient because of the developed software and relatively slow pace of events, and hence, allows easy future extensions and enhancements to the provided data. The pipeline we developed to record, store, retrieve, convert, and annotate videos can easily be adapted to another context: another barn, other behaviors, other animals, and so forth. All the codes necessary to reproduce this pipeline are available in the supplementary material.

Regarding the choice of classes, only a few scenes could not be assigned to any of these classes and mainly only due to difficulties in human interpretation. More critically, we note that future approaches should not prioritize the actions present in the scene, but rather be designed to allow more for flexible analyses of the scenes. For example, drinking and interactions often took place simultaneously but in different parts of the image. The image could be either multi-classified or analyzed in sections. We considered the human class top priority both because we assume that cows act differently when human are present and also due to privacy reasons. It should also be noted that in the area covered in this study, the monitoring of standing-lying behavior or eating patterns was not meaningful, which, nonetheless, are important indicators of a cow's well-being. Thus, extending the camera setup to either cover the whole barn or at least several other well-chosen regions is required.

The deep learning approach we used provided a good fit in terms of accuracy and was feasible in training speed. When analyzing the errors in the predictions, we found that often the scene predicted with the second highest probability would have been the correct. Furthermore, the activation maps are visually quite sufficient. However, the networks did not provide good generalizations outside the time span used to train it. This suggests that the network memorized the training data and closely identical, consecutive frames were present in the training, validation, and test sets.

Previously, aggression has been detected with an accuracy rate of 98% and a recall rate of 98% (Chen et al. 2019); interaction or non-interaction with an accuracy rate of approximately 60% and a recall rate of 100% (Ardo et al. 2017); feeding or non-feeding with an accuracy rate of 92% and a recall rate of 88% (Porto et al. 2015), accuracy rate of 97% (Achour et al. 2020) or 99.4% in a pig study (Alameer et al. 2020); and mounting or non-mounting with an accuracy rate of 91% and a recall rate of 95% (Li et al. 2019). Compared with such prior works, we aimed for a broader set of monitored events in the barn and experimented with two different camera positions and longer data acquisition times, yielding illumination changes. Similar work to ours is reported by Wu et al. (2021), where classification of five different single cow behavioral actions were studied. However, the complexity increases when interactions are studied, as in our work reported in this article. Also, Tsai et al. (2020) studied drinking behavior via

monitoring cow head from above of the throughs. Their approach yielded F1 score of 0.987 and true positive rate of the cow head 0.983, which, however, are not directly compared to our results as we labeled the scene "drinking" only when the cow was engaged in water intake and not just in the vicinity of the through.

In our analyses the temporal domain had excess of data and might benefit from aggregating the sequential frames. For example, studies of pigs by Chen et al. (2020a and 2020b) used one-second episodes and Yang et al. (2019) thirty-second episodes for analysis instead of individual frames. Aggregating frames would both solve the memorizing issue and enable analysis of dynamic events when used together with suitable CNN structures such as long-short-term-memory (see Chen et al. 2020a). With the single frame approach, no distinction was made in the speed of interactions, which we see as a drawback. Dynamic representation of data should include information on the intensity of the events and their directions. In different barn scenarios, such need has already been recognized and started by Guo et al. (2019) to recognize mounting and by Fuentes et al. (2020) who used 15 different behavioral actions of cows – either individual, in a group or pairwise. Their classes are mainly different from our case, thus supplementing each other depending on the data acquisition position within the barn.

Dynamic data representation could also be combined with the above suggested sectioning of the image domain, which would revert the wider scenery towards simultaneous binary analysis, or at least analysis of fewer classes in the whole image. Certain events, like drinking, curiosity and queuing are only possible on certain positions of the image domain. The difference in prediction performance of unbalanced and balanced experiments, where there was a trade-off between accuracy and less bias, might as well be interpreted to suggest implementation of multiclass detection as a bundle of independent CNNs, each detecting only whether an event is present of not.

Availability and variety of data is seen critical for further development of monitoring applications which are suggested to include multiple camera views and be based on standardized data (Li et al. 2021b, Bao and Xie 2022), In this work, we have demonstrated an approach to multiclass scene recognition of cows from video data of their every-day environment and behavior. The provided work and data serve further research as a pre-labeled and pre-trained piece adaptable to new situations.

Data and software availability

We have shared the data used in the article, except that the parts including person were removed for reasons of privacy. Also, the developed Python software is shared. Both the data and the software codes are licensed with Creative Commons 4.0 Attribute license and are available via https://doi.org/10.5281/zenodo.3981400.

Acknowledgments

This work was funded by Finland's Ministry of Education and Culture through the Bioeconomy 4.0 project and by Brazil's Federal Institute of Santa Catarina through the Propicie exchange grant. We thank Joni Kukkamäki and Atte Partanen (HAMK SMART research unit, Häme University of Applied Science) for their help with the data infrastructure and Katri Virtanen (Häme Vocational Institute) and Simo Pärssinen (School of Bioeconomy, Häme University of Applied Sciences) for their help with the practicalities of sensor installation.

References

Achour, B., Belkadi, M., Filali, I., Laghrouche, M. & Lahdir, M. 2020. Image analysis for individual identification and feeding behaviour monitoring of dairy cows based on Convolutional Neural Networks (CNN). Biosystems Engineering 198: 31–49. https://doi.org/10.1016/j.biosystemseng.2020.07.019

Alameer, A., Kyriazakis, I., Dalton, H.A., Miller, A.L. & Bacardit, J. 2020. Automatic recognition of feeding and foraging behaviour in pigs using deep learning. biosystems engineering 197: 91–104. https://doi.org/10.1016/j.biosystemseng.2020.06.013

Anagnostopoulos, A., Barden, M., Tulloch, J., Williams, K., Griffiths, B., Bedford, C. & Oikonomou, G. 2021. A study on the use of thermal imaging as a diagnostic tool for the detection of digital dermatitis in dairy cattle. Journal of dairy science 104: 10194–10202. https://doi.org/10.3168/jds.2021-20178

Ardo, H., Guzhva, O., Nilsson, M. & Herlin, A. 2017. A CNN-based cow interaction watchdog. IET Computer Vision, 12. https://doi.org/10.1049/iet-cvi.2017.0077

Atkinson, G.A., Smith, L.N., Smith, M.L., Reynolds, C.K., Humphries, D.J., Moorby, J.M., Leemans, D.K. & Kingston-Smith, A.H. 2020. A computer vision approach to improving cattle digestive health by the monitoring of faecal samples. Scientific Reports 10: 17557. https://doi.org/10.1038/s41598-020-74511-0

Banhazi, T. & Tscharke, M. 2016. A brief review of the application of machine vision in livestock behaviour analysis. Journal of Agricultural Informatics 7. https://doi.org/10.17700/jai.2016.7.1.279

Bao, J. & Xie, Q. 2022. Artificial intelligence in animal farming: A systematic literature review. Journal of Cleaner Production 331: 129956. https://doi.org/10.1016/j.jclepro.2021.129956

Benitez Pereira, L.S., Koskela, O., Pölönen, I. & Kunttu, I. 2020. Data set of labeled scenes in a barn in front of automatic milking system. https://doi.org/10.5281/zenodo.3981400

Berckmans, D. 2014. Precision livestock farming technologies for welfare management in intensive livestock systems. Revue scientifique et technique (International Office of Epizootics) 33: 189–196. https://doi.org/10.20506/rst.33.1.2273

Boissy, A., Manteuffel, G., Jensen, M., Moe, R., Spruijt, B., Keeling, L., Winckler, C., Forkman, B., Dimitrov, I., Langbein, J., Bakken, M., Veissier, I. & Aubert, A. 2007. Assessment of positive emotions in animals to improve their welfare. Physiology & Behavior, 92:375–97. https://doi.org/10.1016/j.physbeh.2007.02.003

Bradski, G. 2000. The OpenCV Library. Dr. Dobb's Journal of Software Tools.

Buslaev, A., Iglovikov, V.I., Khvedchenya, E., Parinov, A., Druzhinin, M. & Kalinin, A.A. 2020. Albumentations: Fast and flexible image augmentations. Information 11: 125. https://doi.org/10.3390/info11020125

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. Computers and Electronics in Agriculture 166: 105003. https://doi.org/10.1016/j.compag.2019.105003

Chen, C., Zhu, W., Steibel, J., Siegford, J., Han, J. & Norton, T. 2020 a. Recognition of feeding behaviour of pigs and determination of feeding time of each pig by a video-based deep learning method. Computers and Electronics in Agriculture 176: 105642. https://doi.org/10.1016/j.compag.2020.105642

Chen, C., Zhu, W., Oczak, M., Maschat, K., Baumgartner, J., Larsen, M.L. V. & Norton, T. 2020 b. A computer vision approach for recognition of the engagement of pigs with different enrichment objects. Computers and Electronics in Agriculture 175: 105580. https://doi.org/10.1016/j.compag.2020.105580

Chollet, F. 2017. Deep Learning with Python. Manning Publications Co., Greenwich, CT, USA, 1st edition.

Cuthbertson, H., Tarr, G. & González, L.A. 2019. Methodology for data processing and analysis techniques of infrared video thermography used to measure cattle temperature in real time. Computers and Electronics in Agriculture 167: 105019. https://doi.org/10.1016/j.compag.2019.105019

de Sousa, R.V., da Silva Rodrigues, A.V., de Abreu, M.G., Tabile, R.A. & Martello, L.S. 2018. Predictive model based on artificial neural network for assessing beef cattle thermal stress using weather and physiological variables. Computers and electronics in agriculture 144: 37–43. https://doi.org/10.1016/j.compag.2017.11.033

Faceli, K., Lorena, A.C., Gama, J. & Carvalho, A.C.P. de L.F. de 2011. Inteligência articial: uma abordagem de aprendizado de máquina. LTC. 394 p.

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. https://doi.org/10.1016/j.compag.2020.105627

Gardenier, J., Underwood, J. & Clark, C. 2018. Object detection for cattle gait tracking. In International Conference on Robotics and Automation. https://doi.org/10.1109/ICRA.2018.8460523

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

Guo, Y., Zhang, Z., He, D., Niu, J. & Tan, Y. 2019. Detection of cow mounting behavior using region geometry and optical flow characteristics. Computers and Electronics in Agriculture 163: 104828. https://doi.org/10.1016/j.compag.2019.05.037

Guzhva, O., Ardö, H., Nilsson, M., Herlin, A. & Tufvesson, L. 2018. Now you see me: Convolutional neural network based tracker for dairy cows. Frontiers in Robotics and AI 5: 107. https://doi.org/10.3389/frobt.2018.00107

Huang, X., Li, X. & Hu, Z. 2019. Cow tail detection method for body condition score using Faster R-CNN. In 2019 IEEE International Conference on Unmanned Systems and Artificial Intelligence (ICUSAI) IEEE. p. 347–351. https://doi.org/10.1109/ICUSAI47366.2019.9124743

Jiang, B., Song, H. & He, D. 2019 a. Lameness detection of dairy cows based on a double normal background statistical model. Computers and Electronics in Agriculture 158: 140–149. https://doi.org/10.1016/j.compag.2019.01.025

Jiang, B., Wu, Q., Yin, X., Wu, D., Song, H. & He, D. 2019 b. FLYOLOv3 deep learning for key parts of dairy cow body detection. Computers and Electronics in Agriculture 166: 104982. https://doi.org/10.1016/j.compag.2019.104982

Jorquera-Chavez, M., Fuentes, S., Dunshea, F.R., Jongman, E.C. & Warner, R.D. 2019. Computer vision and remote sensing to assess physiological responses of cattle to pre-slaughter stress, and its impact on beef quality: A review. Meat Science 156: 11–22. https://doi.org/10.1016/j.meatsci.2019.05.007

Kang, X., Li, S., Li, Q. & Liu, G. 2022. Dimension-reduced spatiotemporal network for lameness detection in dairy cows. Computers and Electronics in Agriculture 197: 106922. https://doi.org/10.1016/j.compag.2022.106922

Krizhevsky, A., Sutskever, I. & Hinton, G. 2012. Imagenet classiffication with deep convolutional neural networks. Neural Information Processing Systems 25.

Li, D., Chen, Y., Zhang, K. & Li, Z. 2019. Mounting behaviour recognition for pigs based on deep learning. Sensors 19: 4924. https://doi.org/10.3390/s19224924

Li, S., Fu, L., Sun, Y., Mu, Y., Chen, L., Li, J. & Gong, H. 2021 a. Individual dairy cow identification based on lightweight convolutional neural network. Plos one 16: e0260510. https://doi.org/10.1371/journal.pone.0260510

Li, G., Huang, Y., Chen, Z., Chesser, G.D., Purswell, J.L., Linhoss, J. & Zhao, Y. 2021 b. Practices and Applications of Convolutional Neural Network-Based Computer Vision Systems in Animal Farming: A Review. Sensors 21: 1492. https://doi.org/10.3390/s21041492

Liu, H., Reibman, A.R. & Boerman, J.P. 2020. Video analytic system for detecting cow structure. Computers and Electronics in Agriculture 178: 105761. https://doi.org/10.1016/j.compag.2020.105761

Mahmud, M.S., Zahid, A., Das, A.K., Muzammil, M. & Khan, M.U. 2021. A systematic literature review on deep learning applications for precision cattle farming. Computers and Electronics in Agriculture 187: 106313. https://doi.org/10.1016/j.compag.2021.106313

Müller, R. & Schrader, L. 2003. A new method to measure behavioural activity levels in dairy cows. Applied Animal Behaviour Science 83: 247–258. https://doi.org/10.1016/S0168-1591(03)00141-2

Nasirahmadi, A., Edwards, S. & Sturm, B. 2017. Implementation of machine vision for detecting behaviour of cattle and pigs. Livestock Science 202. https://doi.org/10.1016/j.livsci.2017.05.014

OIE 2019. Terrestrial animal health code. Standard, World Organization for Animal Health (OIE) Paris, France.

Porto, S., Arcidiacono, C., Anguzza, U. & Cascone, G. 2015. The automatic detection of dairy cow feeding and standing behaviours in free-stall barns by a computer vision based system. Biosystems Engineering 133. https://doi.org/10.1016/j.biosystemseng.2015.02.012

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. https://doi.org/10.1016/j.compag.2019.104958

Qiao, Y., D. Su, D., Kong, H., Sukkarieh, S., Lomax, S. & Clark, C. 2020. "Data Augmentation for Deep Learning based Cattle Segmentation in Precision Livestock Farming," 2020 IEEE 16th International Conference on Automation Science and Engineering (CASE) p. 979–984. https://doi.org/10.1109/CASE48305.2020.9216758

Russell, S.J. & Norvig, P. 2010. Artificial Intelligence: A Modern Approach. Pearson Prentice Hall, New Jersey.

Schirmann, K., von Keyserlingk, M.A., Weary, D.M., Veira, D.M. & Heuwieser, W. 2009. Validation of a system for monitoring rumination in dairy cows. Journal of Dairy Science 92: 6052–6055. https://doi.org/10.3168/jds.2009-2361

Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D. & Batra, D. 2019. Grad-cam: Visual explanations from deep networks via gradient-based localization. International Journal of Computer Vision 128: 618–626. https://doi.org/10.1007/s11263-019-01228-7

Shao, W., Kawakami, R., Yoshihashi, R., You, S., Kawase, H. & Naemura, T. 2019. Cattle detection and counting in uav images based on convolutional neural networks. International Journal of Remote Sensing 41: 31–52. https://doi.org/10.1080/01431161.2019.1624858

Shen, W., Cheng, F., Zhang, Y., Wei, X., Fu, Q. & Zhang, Y. 2020. Automatic recognition of ingestive-related behaviors of dairy cows based on triaxial acceleration. Information Processing in Agriculture 7: 427–443. https://doi.org/10.1016/j.inpa.2019.10.004

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

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. 2014. Dropout: A simple way to prevent neural networks from overfitting. Journal of Machine Learning Research 15: 1929–1958.

Steeneveld, W. & Hogeveen, H. 2015. Characterization of dutch dairy farms using sensor systems for cow management. Journal of Dairy Science 98: 709–717. https://doi.org/10.3168/jds.2014-8595

Ter-Sarkisov, A., Ross, R. & Kelleher, J. 2017. Bootstrapping labelled dataset construction for cow tracking and behavior analysis. In: 14th Conference on Computer and Robot Vision (CRV). p. 277–284. https://doi.org/10.1109/CRV.2017.25

Ter-Sarkisov, A., Ross, R., Kelleher, J., Earley, B. & Keane, M. 2018. Beef cattle instance segmentation using fully convolutional neural network. arXiv preprint arXiv:1807.01972.

Tsai, Y.-C., Hsu, J.-T., Ding, S.-T., Rustia, D.J.A. & Lin, T.-T. 2020. Assessment of dairy cow heat stress by monitoring drinking behaviour using an embedded imaging system. Biosystems Engineering 199: 97–108. https://doi.org/10.1016/j.biosystemseng.2020.03.013

van den Berg, G., Viazzi, S., Ismayilova, G., Sonoda, T., Oczak, M., Leroy, T., Costa, A., Bahr, C., Guarino, M., Fels, M., Hartung, J., Vranken, E., Berckmans, D., Köfer, J. & Schobesberger, H. 2011. Labelling of video images: the first step to develop an automatic monitoring tool of pig aggression. In: Proceedings of the 15th ISAH Congress.

Wegner, T.N., Schuh, J.D., Nelson, F.E. & Stott, G.H. 1976. Effect of stress on blood leucocyte and milk somatic cell counts in dairy cows. Journal of Dairy Science 59: 949–956. https://doi.org/10.3168/jds.S0022-0302(76)84303-2

Wu, D., Wang, Y., Han, M., Song, L., Shang, Y., Zhang, X. & Song, H. 2021. Using a CNN-LSTM for basic behaviors detection of a single dairy cow in a complex environment. Computers and Electronics in Agriculture 182: 106016. https://doi.org/10.1016/j.compag.2021.106016

Xudong, Z., Xi, K., Ningning, F. & Gang, L. 2020. Automatic recognition of dairy cow mastitis from thermal images by a deep learning detector. Computers and Electronics in Agriculture 178: 105754. https://doi.org/10.1016/j.compag.2020.105754

Yang, A., Huang, H., Yang, X., Li, S., Chen, C., Gan, H. & Xue, Y. 2019. Automated video analysis of sow nursing behavior based on fully convolutional network and oriented optical flow. Computers and Electronics in Agriculture 167: 105048. https://doi.org/10.1016/j.compag.2019.105048

Zin, T.T., Phyo, C.N., Tin, P., Hama, H. & Kobayashi, I. 2018a. Image technology based cow identification system using deep learning. In: International MultiConference of Engineers and Computer Scientists.

Zin, T.T., Tin, P., Kobayashi, I. & Horii, Y. 2018b. An automatic estimation of dairy cow body condition score using analytic geometric image features. In: 7th Global Conference on Consumer Electronics (GCCE). p.775–776. https://doi.org/10.1109/GCCE.2018.8574852