Animal 18 (2024) 101079

Contents lists available at ScienceDirect

Animal

The international journal of animal biosciences

Biometric identification of dairy cows via real-time facial recognition

N. Bergman [a,b], Y. Yitzhaky [a], I. Halachmi [b,*]

[a] School of Electrical and Computer Engineering, Ben-Gurion University of the Negev, 1 Ben Gurion Avenue, P.O.B. 653, Be'er Sheva 8410501, Israel
[b] Precision Livestock Farming (PLF) Laboratory, Institute of Agricultural Engineering, Agricultural Research Organization (A.R.O.) – The Volcani Center, 68 Hamaccabim Road, P.O.B 15159, Rishon Lezion 7505101, Israel

ARTICLE INFO

Article history:
Received 15 February 2023
Revised 2 January 2024
Accepted 11 January 2024
Available online 18 January 2024

Keywords:
Computer vision
Deep learning
Farm management
Feeding behaviour
Real-time monitoring

ABSTRACT

Biometrics methods, which currently identify humans, can potentially identify dairy cows. Given that animal movements cannot be easily controlled, identification accuracy and system robustness are challenging when deploying an animal biometrics recognition system on a real farm. Our proposed method performs multiple-cow face detection and face classification from videos by adjusting recent state-of-the-art deep-learning methods. As part of this study, a system was designed and installed at four meters above a feeding zone at the Volcani Institute's dairy farm. Two datasets were acquired and annotated, one for facial detection and the second for facial classification of 77 cows. We achieved for facial detection a mean average precision (at Intersection over Union of 0.5) of 97.8% using the YOLOv5 algorithm, and facial classification accuracy of 96.3% using a Vision-Transformer model with a unique loss-function borrowed from human facial recognition. Our combined system can process video frames with 10 cows' faces, localize their faces, and correctly classify their identities in less than 20 ms per frame. Thus, up to 50 frames per second video files can be processed with our system in real-time at a dairy farm. Our method efficiently performs real-time facial detection and recognition on multiple cow faces using deep neural networks, achieving a high precision in real-time operation. These qualities can make the proposed system a valuable tool for an automatic biometric cow recognition on farms.

Implications

This study developed a video-based system which performs biometric identification of dairy cows by facial detection and facial recognition of up to 10 cows simultaneously in real time. The method presented provides a way to identify and track each cow individually and monitor its mealtimes and durations. Importantly, it proves that biometric identification can be made based on faces only. In addition, the system can support monitoring of feeding behaviour, which can provide the farmer with practical knowledge about each of the animals being fed, as well as that of all of them as a group, providing another tool to improve the farm's functioning and animals' well-being.

Introduction

Precision livestock farming (**PLF**) can be defined as real-time monitoring technologies aimed at managing individual animals on a farm (Halachmi et al., 2019; Bloch et al., 2019; Geffen et al., 2020; Bezen et al., 2020). At this point, most of the PLF applications in dairy cows are based on monitoring tags attached to the animal (neck, leg, or ear tags). Such examples are ear tag accelerometers (McGowan et al., 2007), neck collars (Roessenet al., 2015; Werner et al., 2017), noseband pressure sensors to record jaw movements (Werner et al., 2017), and core body temperature sensors (Wolfgeret al., 2015; Shahriar et al., 2016). Radio frequency identification-based sensors to locate animals inside buildings are also used (Porto et al., 2014; Pastell et al., 2018).

According to (Kumar et al., 2016), face images of dairy cows are a crucial biometric characteristic for identification due to their unique skin texture and distinct facial features. However, digital facial recognition encounters various well-known challenges, including variations in pose, expression, illumination and aging. These challenges, particularly illumination, pose, and image quality, become evident when capturing facial images of dairy cows.

Thus, more accurate methods are necessary to overcome these challenges and improve real-time animal biometric identification leading to enhanced monitoring of feed intake, resulting in improved farm management decision-making (Shalloo et al., 2004). Animal recognition can also help to manage animals by providing information on animal behaviour, health, welfare, and productivity. For example, it can record activity level, area

[*] Corresponding author.
E-mail address: halachmi@volcani.agri.gov.il (I. Halachmi).

https://doi.org/10.1016/j.animal.2024.101079
1751-7311/© 2024 The Author(s). Published by Elsevier B.V. on behalf of The Animal Consortium.

occupancy, resource use and posture of animals. This can help to detect signs of stress, disease, or injury in the animals and provide feedback for improving their environment and management (Wurtz et al., 2019).

In recent years, computer vision and deep learning methods have made enormous technological leaps (Szegedy et al., 2016). Deep learning, and more specifically, Convolution Neural Networks (**CNNs**) and Transformers, which are specific machine learning methods, are utilized in complicated computer vision tasks, such as detection, classification, recognition, and tracing (Bezen et al., 2020). Due to the quantity of data available (Zhang et al., 2018) and the growth of computer computation capabilities of Graphics Processing Units (**GPUs**) (Tsai et al., 2018), CNNs and Transformers produce increasingly accurate results. They are based on non-linear, end-to-end training, which enables them to learn millions of parameters. Consequently, they require large amounts of diverse and annotated data (Ros et al., 2016). The fields of precision agriculture, agricultural robotics, PLF, and others aimed at developing agricultural applications have seen a rise in interest in implementing deep learning methods, especially CNN-based algorithms (Zheng et al., 2018; Yang et al., 2018; Wang et al., 2018a; 2018b; Alvarez et al., 2018; Pu et al., 2018; Tian et al., 2019; Xu et al., 2020; Denholm et al., 2020). Using CNN algorithms and facial images, biometric identification of dairy cows was performed in order to recognize and identify each animal based on their individual facial features (Yao et al., 2019; Yang et al., 2019; Qiao et al., 2019; Wang et al., 2020; Xu et al., 2022).

However, the methods cited above required manually recording images of cow faces. The studies mentioned above are based on data captured by front-view hand-held cameras, making the methods difficult to apply in a commercial dairy farm. Furthermore, the researchers did not always optimize their algorithms to run at real-time, further limiting practical application.

The aim of this study is to develop and validate a system which detects and identifies multiple cows simultaneously from a video camera, based only on images of their faces, on a commercial dairy farm.

We hypothesize that machine-vision and deep-learning methods have the potential to accurately identify cows solely based on their facial characteristics in real-world commercial dairy farm settings, achieving a high level of accuracy. The method proposed here utilizes a system that automatically captures images of the faces of individual cows or of multiple cows; the system was installed in the feeding areas of a dairy farm. Using an efficient lightweight algorithm in inference is the key for later implementations of the system in real dairy farms.

Material and methods

Animals and housing

We built and installed our prototype system in the Israeli Agricultural Research Organization (**ARO**) dairy farm with the two cameras installed in two separate feeding areas. The system (Fig. 1) consists of a unique metal structure which is constructed from three 1.6 m beams and a rounded rail 12 m long, two IP HIKVISION DS2CD2646G1-IZS cameras which can capture and record four megapixel videos, a wireless router, a power-over-ethernet switch, an uninterruptible power supply unit and a Latte Panda controller; the controller and router are in a closed plastic box to protect them from dirt and damage. Both cameras are connected through Ethernet cables. The cameras were positioned 4 m above the ground, precisely located in the middle of the edge of each feeding area. The feeding area is a square measuring 6 m in length and 1 m in width and serves as a designated space where cows gather to feed. Each camera is able to capture an area of 1 × 6 m, which can accommodate up to 10 cows feeding simultaneously. We used two cameras to cover two adjacent feeding areas resulting in a total rectangular feeding area of 1 × 12 m. We recorded 20 days in total, in two periods, the first period from December 2nd to December 12th 2020, and the second from April 4th to April 14th 2021. Through an inspection of the recorded videos of the dairy cows' behaviour before the system installation and after its removal, apparently cows' behaviour was not changed. The dairy cows exhibited minimal awareness of its existence and functionality. A total of 77 cows were observed during both recording periods.

Facial detection and facial recognition system workflow

The computerized system (Fig. 2) consists of two main algorithms that operate sequentially, (a) a facial detection and localization, based on a CNN which detects multiple cows' faces from an image and automatically crops them, and (b) a facial recognition deep neural network (**DNN**), which classifies cow identities based on facial biometrics which are represented as output feature vectors. In the initial step, faces are detected simultaneously from every video frame by the facial-detection CNN. Subsequently, these detected faces are processed and inputted into the facial-recognition DNN, enabling classification of all the identified cow faces within the frame as a single batch. A batch refers to a set of input data samples that are processed together in a single forward and backward pass through the neural network. The system operates by independently processing each video frame and employs the trained DNNs of both facial detection and facial recognition tasks. This selection of algorithms ensures efficient real-time execution.

Data collection

Cow facial detection dataset

We designed our system to process real-time **RGB** (Red, Green, and Blue) video files. As an initial step, to train the system, we created a cow facial detection dataset based on annotated bounding-boxes in YOLO format (Redmon et al., 2016). We used"Labelimg", which is a tool (Lin, 2015) for annotating faces manually. Our facial detection dataset is solely composed of cow faces and focuses exclusively on a single category, which is a cow face. Through training the facial detection CNN with this dataset, we can efficiently detect a diverse range of cow faces within a frame. We annotated 2 164 cows' faces in bounding boxes, from 500 images. The entire dataset is based on still frames taken from videos acquired during both recording periods. We used this dataset to train our cow facial detector, which can automatically detect, localize, and crop all cow faces from each frame.

Cow facial recognition dataset

We created another dataset of the 77 cows' identities based on the facial images recorded, which we called COW77. The dataset contains 7 032 cropped single facial images from our recorded video files. A sample of 12 face identities is shown in Fig. 3. We semi-automatically annotated our images by applying our pre-trained cow facial-detection algorithm to the video frames, to save time.

Algorithms and models

Facial detection and localization

Our facial detection algorithm is based on the YOLOv5 algorithm (Jocher, 2020). We also used YOLOv5n, YOLOv5s, and YOLOv5m which refer to Nano, Small and Medium variations, men-

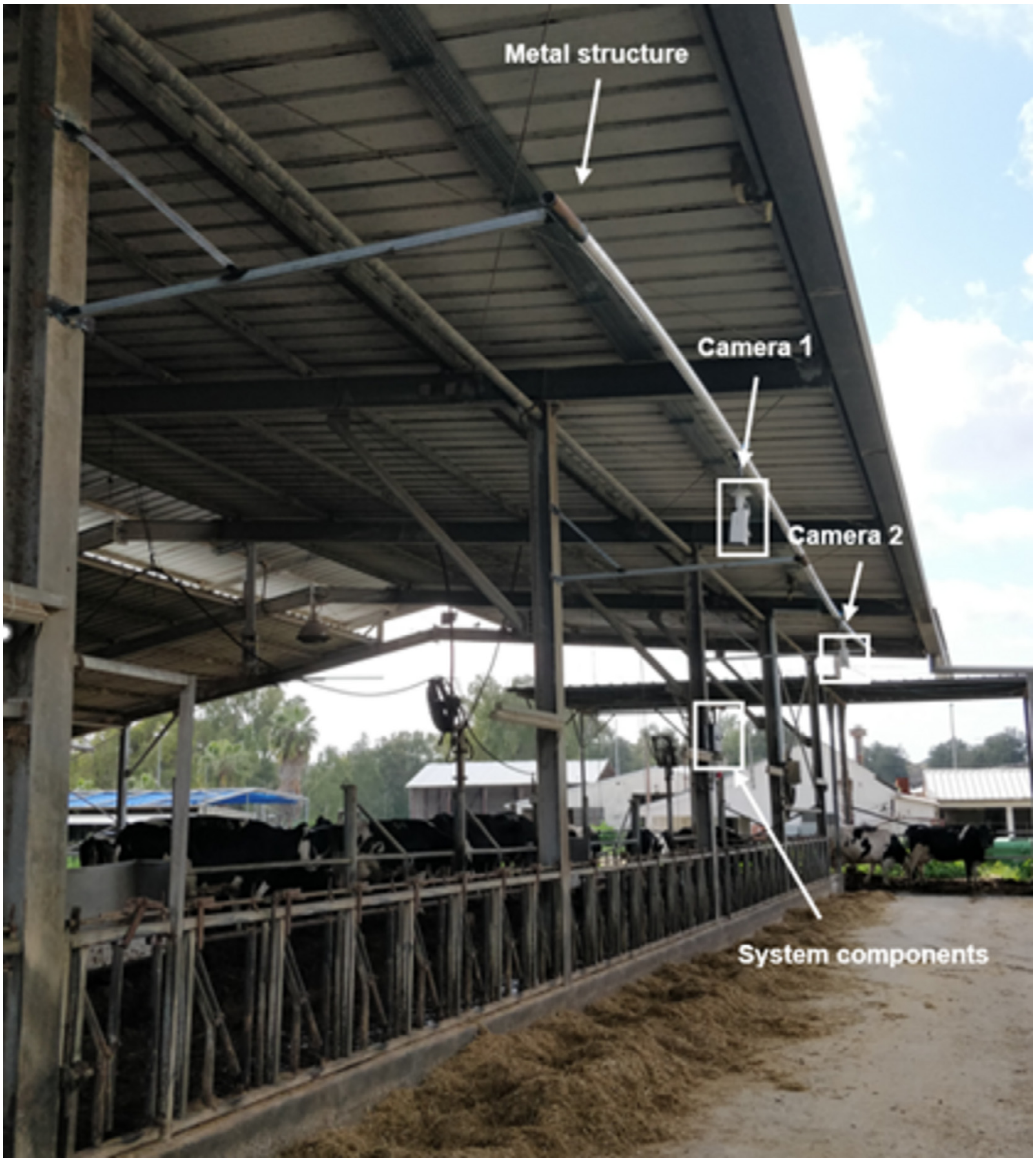

Fig. 1. Our specially built system installed in the Volcani dairy farm at a height of 4 meters above the feeding zone. Using two video cameras, we captured footage of 77 Holstein cows during their feeding times in two adjacent 1 × 6 m rectangular areas. Each camera can record the faces of up to 10 cows simultaneously. The system components include a wireless router, a Power-over-Ethernet switch, an uninterruptible power supply unit and a Latte-Panda controller.

tioned in the results section. The various formats are distinguished from one another by depth and scale multiples (Tan et al., 2020). We selected YOLOv5 as our facial-detection algorithm, due to its efficiency in real-time video-based systems (Solawetz, 2021), and because it is a single-stage object detector (Zaidi et al., 2022).

Facial recognition feature extraction deep neural network architecture

Our selected facial feature extraction DNN encodes the facial input images into vectors for each model we examined, where *d* is the size of a DNN's facial feature vector, which is set to 512, as established by (Wen et al., 2016; Liu et al., 2017; Zhang et al., 2017). That is, each facial image is encoded, and then converted into a 77-sized probability vector for each cow identity; the predicted identity will be the one with the highest probability. We compared nine DNN architectures for this feature extraction, at two scales, based on their number of parameters, floating points operations per second, and inference speed. We divided them into Scale A, which are small models, and Scale B, which are tiny models. Among them, we compared Vision Transformer (**ViT**) (Dosovitskiy et al., 2020) and Visformer (Chen et al., 2021), which are Transformer-based methods. The other architectures are state-of-the-art feature encoder CNNs. Among them, we compared ResNet (He et al., 2016), EfficientNet (Tian et al., 2019), DenseNet (Huang et al., 2017) and MobileNet (Howard et al., 2017). Transformers architectures are based on multi-headed self-attention mechanisms, which can capture global information by interactions between sequences, converting image patches (square tiles) into vector sequences, and are then processed from patches of the image in computer-vision tasks (Khan et al., 2021).

Transformers lack some of the inductive biases inherent in CNNs, such as translation equivariance and locality (Dosovitskiy et al., 2020). However, if the Transformer models are pretrained on larger datasets and then fine-tuned to the target task with a small dataset (Devlin et al., 2019), such as in our COW77 dataset, they may achieve state-of-the-art results which are better than the aforementioned CNNs. In our case, we used

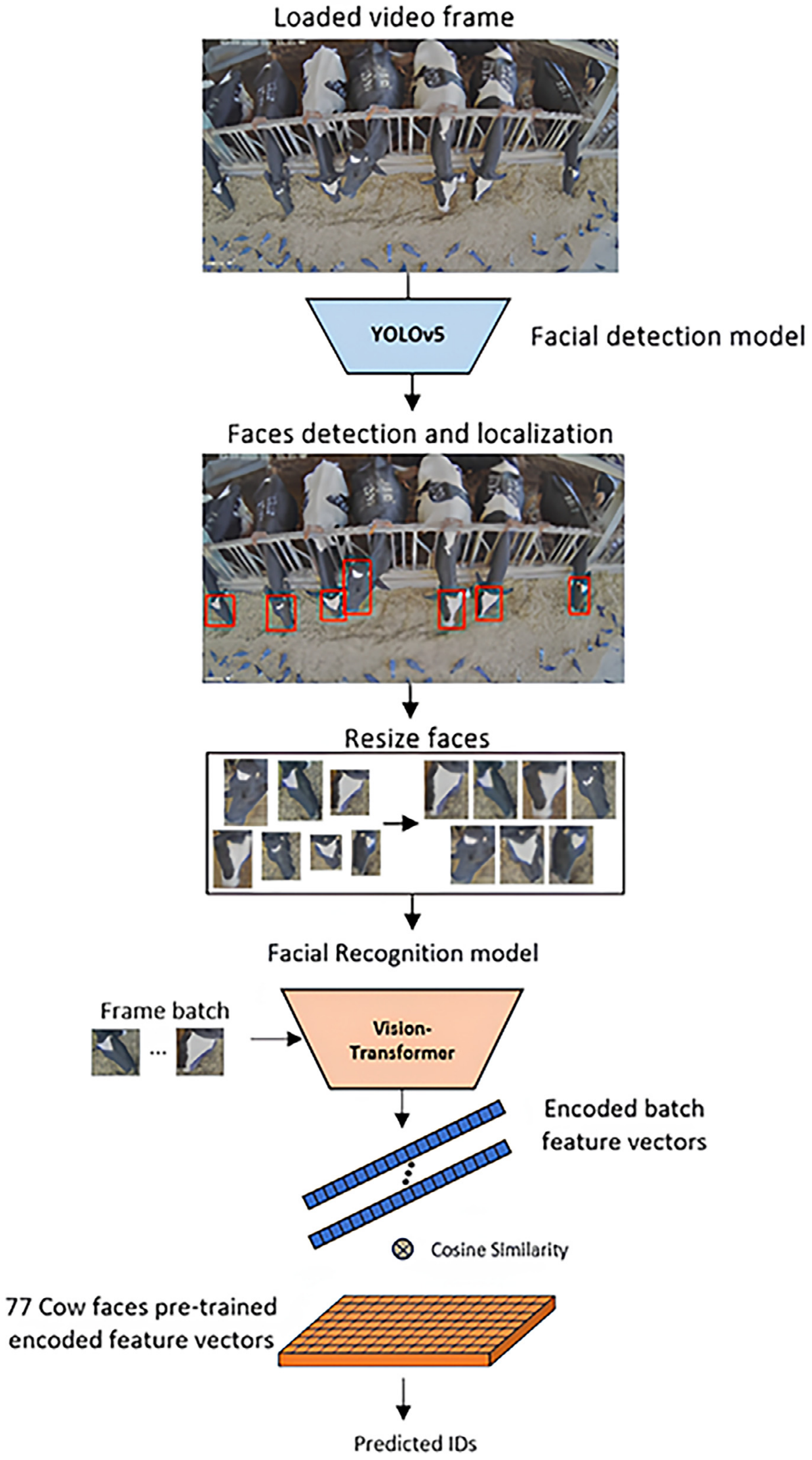

Fig. 2. Cow facial detection and facial recognition system scheme. Each frame is fed into a facial detection convolutional neural network and then all of the detected cows' faces are fed in a single batch into the facial recognition Vision-Transformer deep neural network to identify all faces simultaneously.

transfer-learning for ViT-small and ViT-tiny models with pre-trained weights from the ImageNet dataset (Deng et al., 2010; Ridnik et al., 2021). For our cow facial classification task, these Transformer models proved to be more accurate than other models. Both ViT and CNN models received 8-bit RGB facial images with input size of 224 × 224 × 3 pixels and processed them into a 512-dimensional features vector, which was fed to the loss layer function during training.

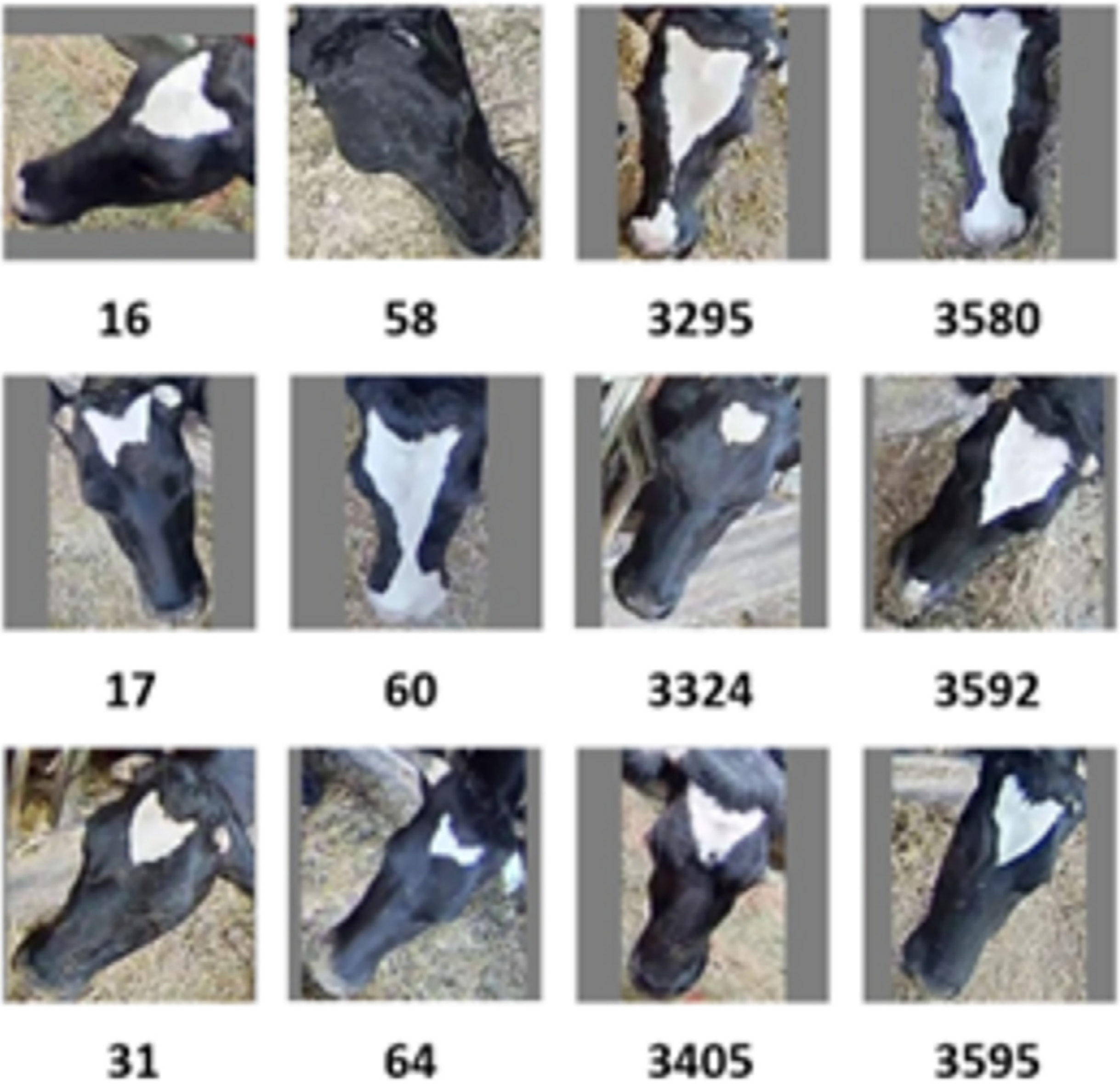

Fig. 3. Gallery of a sample of 12 facial images, which were identified by the system. A total of 7 032 facial images were captured from a variety of facial angles, distances from the cameras, with various occlusions, and illumination conditions. The distribution of face images varies among cow identities, ranging 24–167 images per cow identity.

Facial recognition loss function

During the facial recognition training phase, we used a unique loss function, Additive Angular Margin Loss (ArcFace) (Deng et al., 2022) which is a state-of-the-art technique in human facial recognition. The idea behind this loss-function is to train the model to better recognize each cow's face by increasing the distance between the decision boundaries of features' vectors from identities, and at the same time decrease the distance between facial vectors from the same identity, i.e., increase inter-class discrepancy and improve intra-class compactness using the same mathematical formula (Deng et al., 2022). The facial feature vector output is defined as $x_i \in \mathbb{R}^{512}$, where i is the ground-truth identity.

The representing center of each face identity which comes before the final decision layer is defined as $W_j \in \mathbb{R}^{77\times512}, j = [1, 2, .., 77]$. In ArcFace loss layer, both x_i and W_j are L2-normalized at the outset. We then calculate the cosine similarity index by simply multiplying them, i.e., $\frac{x_i}{||x_i||} \cdot \frac{W_j}{||W||_j}$. This is the projection of the normalized feature vector x_i upon the normalized centers matrix W, which in turn will give $cos\theta_j \epsilon \mathbb{R}^{77}$. θ_j represents the angles vector between face feature vector x_i to each identity center W_j. We then retrieved all angles θ_j from $cos\theta_j$ using an *arccos* operator. We then added a marginal hyper-parameter m to the θ of ground truth identity only, i.e., $\cos(\theta_{y_i} + m)$ while the remaining θ_j angles when $j \neq i$, remain θ_j. We then reapplied the cosine operator, thereby obtaining a modified cosine similarities vector, which we modified for ground truth identity only. The addition of m is in radians and therefore increased the angular distance between the facial features vectors and their ground truth center features vector during training of the network. The m thus acts as a penalization which in turn increases the angular distance between the decision boundaries of each ground truth identity. We then multiplied the modified $cos\theta_j$ vector by a spherical hyper-parameter s, which is the radius of the sphere manifold. The multiplication of s increases the angular distance between facial the features vectors to the center features vector from the same identity during the training. This penalization decreases cosine similarity between the facial image feature vectors to its ground truth center features vector. We then apply a regular SoftMax operation to convert the cosine similarity indices vector into a probability distribution and performed Categorical Cross-Entropy (**CCE**) loss. We also compared another loss function, CosFace loss (Wang et al., 2018a; 2018b) which performs $(\cos(\theta_{y_i}) - m)$ and multiplies in s. The angular-based loss functions described here increased facial recognition accuracy in theory; their mathematical formulae are described in (1)–(3).

$$L_{CCE} = -\frac{1}{N}\sum_{i=1}^{N} log \frac{e^{W_{y_i}^{T}x_i}}{\sum_{j=1}^{n} e^{W_j^{T}x_i}} \tag{1}$$

$$L_{ArcFace} = -\frac{1}{N}\sum_{i=1}^{N} log \frac{e^{s(\cos(\theta_{y_i}+m))}}{e^{s(\cos(\theta_{y_i}+m))} + \sum_{j=1,j\neq y_i}^{n} e^{s\cos\theta_j}} \tag{2}$$

$$L_{CosFace} = -\frac{1}{N}\sum_{i=1}^{N} log \frac{e^{s(\cos(\theta_{y_i})-m)}}{e^{s(\cos(\theta_{y_i})-m)} + \sum_{j=1,j\neq y_i}^{n} e^{s\cos\theta_j}} \tag{3}$$

where N is the batch size. We trained and compared ArcFace and CosFace under three marginal hyperparameter values, $m = 0.1, 0.25, 0.5$, and a single value for $s = 10.0$ in our experiments, as described in the Results section. We also compared the results to regular SoftMax CCE loss. Both methods improved the accuracy test results, while ArcFace was proven to have better, and more stable results compared to those of CosFace.

Intermediate processing between facial-detection and facial-recognition stages

The output of the facial detection system, which is the input of the facial recognition system, includes the detected and cropped cows' facial images, which vary in size and aspect ratios (i.e., the proportion between its width and height), due to the distances from the cameras and the proportions of the detected. To find the best facial input format for the recognition system, we examined two processing methods: (a) maintaining facial aspect ratio of the facial images, and (b) interpolating the facial images without retaining the aspect ratio. Both techniques use bilinear interpolations of the facial images to increase the inference speed. At the end of this process, both processing methods will take captured face images in any given size from the facial detection CNN and convert them into a 224 × 224 × 3 pixels image.

Evaluation metrics

Facial detection and localization evaluation metrics

To assess the accuracy of the facial detection CNN results, we employed two metrics (explained below): mean Average Precision (**mAP**) at Intersection over Union (**IoU**) equals to 0.5 and mAP at IoU equals to 0.5:0.95 (average of mAP for each IoU threshold in increments of 0.05).

mAP is a widely used metric in computer vision that evaluates the performance of object detection algorithms. It considers both accuracy and precision in detecting objects within an image. By comparing predicted and ground truth bounding boxes and measuring the overlap using IoU, mAP calculates precision and recall values. These values are averaged for each object class, and the mean average precision across all classes provides an overall assessment of the algorithm's performance in accurately and completely identifying objects in images. IoU compares the overlap between the predicted facial regions of the cow and the ground truth (actual) facial regions of the cow (pre-annotated). For the first metric, mAP at IoU = 0.5, we evaluated the network's performance by considering a minimum overlap of 50% between the predicted and actual facial regions. The second metric, mAP at IoU = 0.5:0.95, calculates the average precision for a range of IoU thresholds from 0.5 to 0.95, with increments of 0.05. This provides a more comprehensive evaluation of the network's performance across different levels of overlap between predicted and actual facial regions (Padilla et al., 2021).

Precision is the ability of a model to identify only the correct faces, and Recall is the ability of a model to find all ground truth bounding boxes of faces (Padilla et al., 2021). They are defined in formulae (4):

$$\begin{aligned} Precision &= \tfrac{True\ \ Positives}{True\ \ Positives + False\ \ Positives} \\ Recall &= \tfrac{True\ \ Positives}{True\ \ Positives + False\ \ Negatives} \end{aligned} \tag{4}$$

The precision-recall curve is obtained by plotting the model's precision and recall values as a function of the model's confidence score threshold, while the average precision for each class is obtained by interpolating the precision at each level, taking the maximum precision whose value is greater or equal to R_{n+1}.R_n is the $n-th$ recall value, where n is the index for each recall level corresponding to a certain threshold and R_{n+1} is the recall level at the next threshold. P_n is the precision at the $n-th$ recall value. The Average Precision and mAP are defined in formulae (5) and (6).

$$\begin{aligned} AP &= \textstyle\sum_n (R_{n+1} - R_n)P_n(R_{n+1}) \\ P_n(R_{n+1}) &= max_{\tilde{R}:\tilde{R}\geq R_{n+1}} P(\tilde{R}) \end{aligned} \tag{5}$$

$$mAP = \frac{1}{N_c}\sum_{i=1}^{N} AP_i \tag{6}$$

N_c is the total number of classes being evaluated.

Face recognition evaluation metrics

We use accuracy, balanced accuracy, and F1-score as our lead metrics to examine the performance of our face recognition models. True Positives and True Negatives are the elements correctly classified by the model. False Positive are the elements that were labelled as positive by the model, but are actually incorrect, and False Negatives are elements that were labelled as negative by the model, but are actually correct (Grandini et al., 2020). Accuracy uses the sum of TP and TN elements as the numerator, and the sum of all the entries of the confusion matrix as the denominator:

$$Accuracy = \frac{True\ \ Postives + True\ \ Negatives}{True\ \ Positives + True\ \ Negatives + False\ \ Positives + False\ \ Negatives} \tag{7}$$

The Balanced Accuracy is an average of recalls. First, we evaluated the recall for each class, and then, we averaged the values in order to obtain the Balanced Accuracy score.

$$Blanced\ \ Accuracy = \frac{\frac{TruePositives}{Total_{row_i}} + \sum_{j=m,j\neq i}^{M} \frac{TrueNegatives}{Total_{rowm}}}{M} \tag{8}$$

In a multi-class confusion-matrix, M is the number of rows, i.e., categories; $Total_{rowi}$ is the sum of the false negatives; and a single true positive value for the ground-truth row i. $Total_{rowm}$ is the sum of the true negatives and a single false postive value for the remaining categories. In fact, the principal difference between Balanced Accuracy and Accuracy emerges when the dataset shows an unbalanced distribution for the classes. Balanced Accuracy consists of the arithmetical mean of the recall of each class. Therefore, it is balanced because every class has the same weight and the same importance. Consequently, smaller classes eventually have a greater proportional influence on the formula. The formula of the F1-score can be interpreted as a weighted average between Precision and Recall, where the F1-score reaches its best value at 1 and its worst at 0.

$$\text{F1 score} = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \quad (9)$$

Training settings and hyper-parameters

The training processes of both face detection and recognition models were done with a single Nvidia A100 GPU. For the face recognition models training, we used train and validation batch sizes of 64, SGD optimizer with momentum ($\gamma = 0.9$), and the weight decay λ was set to $1 \cdot 10^{-4}$. The initial learning rate was set to 0.05 across models with a step learning rate scheduler of 3% decay in learning rate value every three epochs. The total number of training epochs of the facial recognition DNN was set to 150, and COW77 was randomly sampled in each training session, while keeping a ratio of 65, 17.5, and 17.5% between the training, validation, and test sets, respectively. The same original COW77 dataset distribution between categories was kept in the training-validation-test sets after splitting. Each model was trained four times. Each of the results that appear here is an average of the test-set results from these four training sessions. Facial detection models were trained for 1 500 epochs using the same hardware and predefined hyperparameters of the YOLOv5 official release.

Results

The results of five comparisons are described below: (a) The first compares the facial detection models; (b) the second assesses the intermediate processing of the detected cow faces between facial-detection and facial-recognition stages; (c) the third is the facial recognition models comparison; (d) the fourth is a loss-functions comparison of the selected facial recognition model; (e) the last is the entire system inference time comparison per frame with 10 cow faces.

Comparison of facial detection and localization models

Results of the performances of the facial detection models (YOLOv5-Nano, YOLOv5-Small and YOLOv5-Medium) are shown in Table 1. Each of the models was trained for 1 500 epochs. We achieved with these models mAP rates at IoU threshold of 0.5 of 96.3, 97.4 and 97.9%, respectively. The precision rates were 93.4, 94.5 and 95.9% and recall rates were 89.2, 93.2 and 93.2%, respectively, for YOLOv5-Nano, YOLOv5-Small and YOLOv5-Medium. The training curves of the YOLOv5-Small facial detection model are presented in Fig. 4.

Intermediate processing: the effect aspect ratio on the facial recognition accuracy

Before employing the face-recognition stage, we processed the input images into blocks of 224 × 224 × 3 pixels (i.e., each channel, red, green and blue, is in the size of 224 × 224 pixels). Utilizing the ViT-small model, trained with ArcFace layer incorporating parameters m = 0.25 and s = 10.0, we attained accuracy, balanced accuracy, and F1-score results of 96.3, 96.2, and 96.3% respectively. When we kept the original aspect-ratio of each detected face image by padding the shortest dimension with intensity level values of 0, we achieved accuracy, balanced accuracy, and F1-score of 94.4, 94.2 and 94.4%. The differences between these processing methods in accuracy, balanced accuracy and F1-score are 1.9, 2.0, and 1.9%, respectively. Therefore, we recommend resizing the detected and cropped face images into a square 224 × 224 × 3 pixels uniform RGB image size, without keeping the original face aspect ratio of each image.

Facial-recognition model comparison

ViT-small from Scale A achieved accuracy of 96.3%, balanced accuracy of 96.2% and F1-score of 96.3%. ViT-tiny from Scale B achieved accuracy of 93.9%, balanced accuracy of 93.8% and F1-score of 94.0%, both utilizing the COW77 test-set data (Table 2). Transformer-based architectures achieved better results than did other known convolutional state-of-the-art neural networks. In Scale A, the second-best architecture was achieved by ResNet50, with a 95.8% accuracy, 95.9% balanced accuracy, and F1-score of 95.7%. In Scale B, the second-best architecture was achieved by ResNet18, with 93.6% accuracy, a 93.3% balanced accuracy, and a 93.6% F1-score. The training curves of the ViT-small facial recognition model are presented in Fig. 5.

Facial recognition loss-function study

We compared three loss functions for cow facial-recognition and tested their results on our COW77 dataset. The loss functions are ArcFace, CosFace and regular Softmax with CCE loss. We tested and compared both ArcFace and CosFace for m values of $0.1[rad], 0.25[rad], 0.5[rad]$; s was set to 10.0 (Table 3). ArcFace loss appears to be better than that of CosFace and regular Softmax CCE losses with accuracies of 96.3, 96.3 and 96.4%, balanced accuracies of 96.1, 96.2 and 96.1% and F1-scores of 96.3, 96.3, 96.3% respectively. We noticed that when training with $m = 0.1[rad]$, the training and validation sets appear to be more accurate by 0.6, 0.7, and 0.3% respectively than were the test set results. With $m = 0.25[rad]$, the training and validation set appear to be equal in accuracy as those of the test-set, and in $m = 0.5[rad]$, the training and validation sets tend to be less accurate by 0.3, 0.2, 0.4% respectively than those of the test-set. The accuracy of CosFace losses for $m = 0.1[rad]$ = 0.1 were lower by 0.9, 1.0, and 0.9% for accuracy, balanced-accuracy and F1-score respectively compared with those of ArcFace's losses. Overall, we recommend an ArcFace loss with a marginal hyperparameter m value of 0.25, which achieved the highest results in the validation and test sets.

Table 1
Facial detection model performances, speed of inference and parameter sizes. The detection convolutional neural network was trained with a total of 2 164 bounding boxes framing Holstein cows' faces across 500 frames. The results on the test set that included 325 other face images are derived from bounding boxes representing faces with diverse angles, sizes, illumination conditions, and partial occlusions.

Model	mAP[1]	Precision[2]	Recall[2]	Inference speed[3]	Parameters[4]
YOLOv5-Nano	96.3	93.4	89.2	9.0	7.26
YOLOv5-Small	97.4	94.5	93.2	11.6	21.2
YOLOv5-Medium	97.9	95.9	93.2	19.6	46.5

Abbreviation: mAP = mean Average Precision;
[1] mAP is mean average precision at intersection over the union threshold of 0.5, measured in percentage (%).
[2] Precision is the model's ability to identify only the correct faces, and Recall is the model's ability to find all ground truth bounding boxes of faces.
[3] Inference speed in ms per frame for facial detection when the feeding zone is full (10 cows).
[4] Parameter numbers are in millions (lower is more efficient).

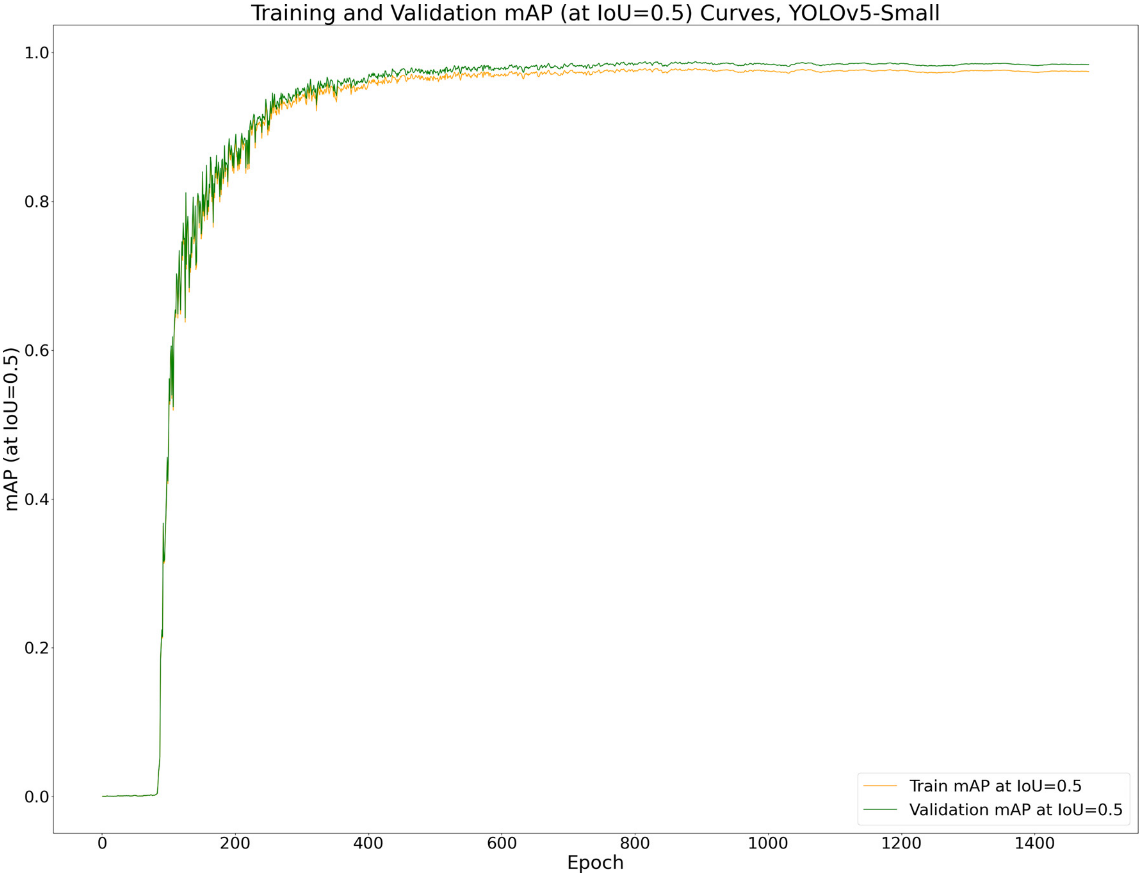

Fig. 4. Mean average precision at intersection over union of 0.5 using YOLOv5-small convolutional neural network during the face detection training process with regard to the number of epochs. Out of 2 164 bounding boxes in the cow face detection dataset, 1 514 bounding boxes were used in the training set and 325 in the validation set.

Table 2
Comparison of facial recognition models: evaluating performance and parameter size for cow facial recognition models at scale A and scale B model sizes. Overall, the recognition deep neural network was trained with 7 032 face images representing 77 individual cows. The reported results are derived from 1 230 test-set face images, maintaining the original distribution.

Model scale	Classification model	Accuracy[1]	Balanced-accuracy[1]	F1-Score[1]	Parameters[2]
A	ViT-small	96.3	96.2	96.3	22.05
	ResNet50	95.8	95.9	95.7	25.56
	Visformer small	94.8	94.9	94.9	40.22
	DenseNet121	94.7	94.8	94.8	7.98
	EfficientNetB3	94.5	94.3	94.6	12.23
B	ViT-tiny	93.9	93.8	94.0	5.72
	ResNet18	93.6	93.3	93.6	11.69
	MobileNetV2	93.2	93.2	93.3	3.5
	MnasNet	92.8	92.7	92.8	4.38

Scale A = middle-to-small scale facial recognition models; Scale B = tiny scale facial recognition models.
[1] Accuracy, Balanced-accuracy and F1-score appear in percentages.
[2] Parameter numbers are in millions (lower is more efficient).

Comparison of inference time for the entire system

Inference speeds were tested on single RTX4000 GPU, YOLOv5-Medium, YOLOv5-Small and YOLOv5-Nano for facial-detection task run in 8.6, 11.4 and 19.4 ms per frame with a fully occupied feeding zone with 10 faces (Table 4). The ViT-small and ViT-tiny models run at 14.2 and 8.0 ms per frame, while the ResNet50 and ResNet18 models run at 7.3 and 3.8 ms per frame which makes the ResNets very efficient. Using YOLOv5-Small will process the frame in 2.8 ms more than YOLOv5-Nano does and will increase the mean average precision by 1.1%. Coupling it with ResNet50 results in a total frame processing time of 18.7 ms per frame. As noted above, ViT-small and ViT-tiny achieved the best Scale A and Scale B results, and therefore, we recommend using them. Fig. 6 presents an example of the output of the system in each frame, specifically detailing the accurate positioning of each identified individual dairy cow and its biometric facial recognition.

Discussion

In this study, we developed a video-based system capable of real-time facial detection, localization, and biometric identification for dairy cows, solely based on their faces.

For the cow face detection stage in our system, we used three model versions of the YOLOv5 family, which vary in sizes and per-

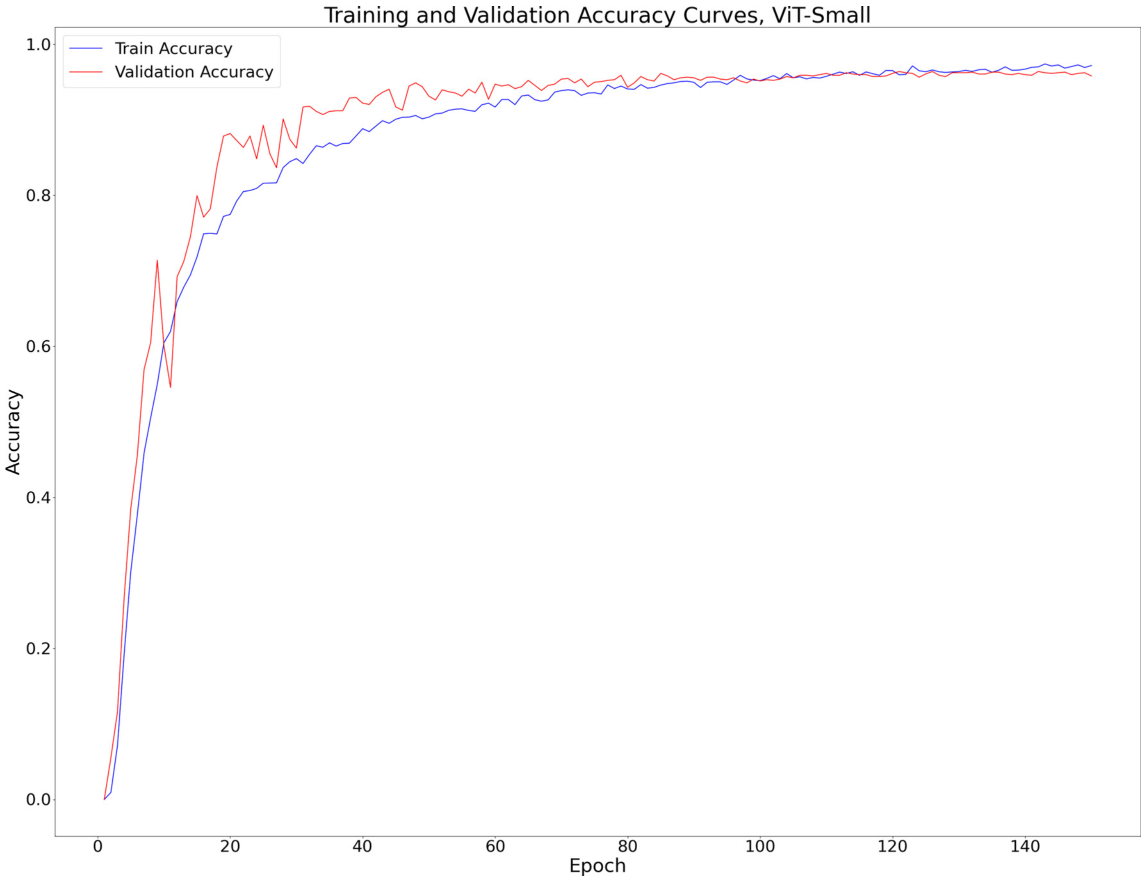

Fig. 5. Recognition accuracy of ViT-small deep neural network during the training process with regard to the number of epochs. Out of 7 032 face images in the cow face recognition dataset, 4 572 face images were used in the training set and 1 230 in the validation set.

Table 3
Comparison of the loss layers in the facial recognition step: ArcFace layer, CosFace layer, and regular Softmax layer, applied with the ViT-small model and Categorical Cross-Entropy as a cost function for identifying cows' faces. Training with an ArcFace loss layer demonstrates an improvement in all metrics compared to using regular Softmax layer or CosFace layer with Categorical Cross-Entropy cost function.

Loss type	s[1]	m[2]	Accuracy	Balanced accuracy	F1-Score
ArcFace	10.0	0.1	96.3	96.1	96.3
		0.25	96.3	96.2	96.3
		0.5	96.4	96.4	96.3
CosFace	10.0	0.1	95.4	95.1	95.4
		0.25	95.7	95.4	95.7
		0.5	96.5	96.3	96.5
Softmax	–	–	95.7	95.4	95.7

ArcFace cost function (Deng et al., 2022); SphereFace cost function (Liu et al., 2017).

[1] s is scale parameter (loss multiplicative). Controls the degree features are stretched to increase class separability on a hypersphere manifold.

[2] m is margin parameter (loss additive). It sets the size of the angular margin between class boundaries.

formance levels. YOLOv5-Nano is the smallest and fastest, while YOLOv5-Medium is six times bigger in terms of parameter number than YOLOv5-Nano and two times bigger than YOLOv5-Small. Based on our findings, the difference in inference speed between YOLOv5-Nano and YOLOv5-Small is about 2 ms per frame and therefore is neglectable, while YOLOv5-Medium is almost two times slower than YOLOv5-Small.

Regarding performance metrics, in mAP at IoU = 0.5, YOLOv5-Medium outperformed YOLOv5-Small by 0.5%, and YOLOv5-Small outperformed YOLOv5-Nano by 1.1%. Precision showed a similar pattern, with YOLOv5-Medium surpassing YOLOv5-Small by 1.4%, and YOLOv5-Small surpassing YOLOv5-Nano by 1.1%. In terms of recall, both YOLOv5-Medium and YOLOv5-Small achieved a recall of 93.2%, while YOLOv5-Nano lagged by 3%. This indicates a notable decrease in cow face detection compared to YOLOv5-Small and YOLOv5-Medium. Although YOLOv5-Nano demonstrated excellent memory efficiency, it did not exhibit the desired inference speed.

Considering mAP, precision, recall, parameter count, and inference speed, we believe that YOLOv5-Small offers the best trade-off. Therefore, when implementing a system in a real dairy farm, we will opt for this model.

The other part of the method deals with cow facial recognition. This part consists of nine algorithms tested at two model scale levels. In Scale A, which includes small-to-middle size models, the study examines ViT-small and Visformer-Small from the Vision-Transformer family, along with ResNet50, DenseNet121, and EfficientNetB3 from the Convolution Neural Networks family.

ViTs and CNNs are two distinct types of models. Recent research has shown that ViT can achieve comparable or superior performance to CNNs in image classification tasks (Dosovitskiy et al.,

Table 4
Comparison of the real-time latency per frame of different combinations of three facial detection models and four facial recognition models in a fully occupied feeding zone with ten cows.

Facial detection model	Facial recognition model	Facial detection inference time[1]	Facial recognition inference time	Total time[2]
YOLOv5-Nano	ViT-tiny	8.6	7.9	16.3
	ViT-small		14.3	22.7
	ResNet18		3.8	12.2
	ResNet50		7.4	15.8
YOLOv5-Small	ViT-tiny	11.4	7.9	19.3
	ViT-small		14.5	25.9
	ResNet18		3.8	15.2
	ResNet50		7.2	18.6
YOLOv5-Medium	ViT-tiny	19.4	8.1	27.5
	ViT-small		13.8	33.2
	ResNet18		3.8	23.2
	ResNet50		7.2	26.6

[1] Time refers to ms per frame, calculated using the selected model on a fully occupied feeding zone with 10 cows' faces.
[2] Total time refers to complete inference time in ms per frame, which includes facial detection of 10 cows simultaneously in a full feeding zone, processing the images, and facial recognition.

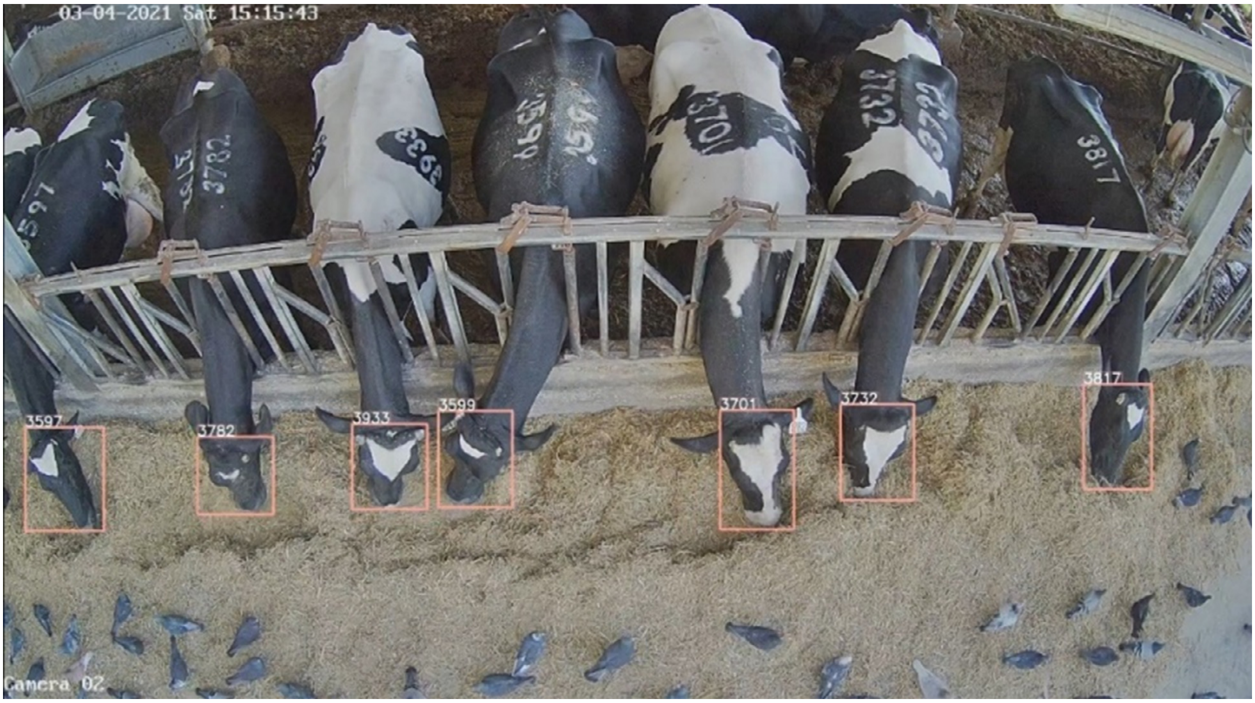

Fig. 6. An example of the system output for a single video frame. The cows' faces were detected by the YOLOv5-small convolutional neural network and classified by the ViT-small deep neural network, in real time. The detection and recognition procedures operate consecutively, taking together 25.9 ms per frame for a fully occupied frame containing 10 cows simultaneously.

2020). We tested this idea using both ViT and CNN-based DNNs and found that it holds true for our data. When comparing the results, there is a notable decrease in performance between Scale A and Scale B. However, both variants of the ViT models, in Scale A and Scale B, achieved the best outcomes. In Scale A, the ViT-small model exhibited a 0.5% higher accuracy and a 0.6% higher F1-score compared to the second-best model, ResNet50, which is a CNN-based model. ResNet50 and ViT-small have a similar number of parameters. In terms of inference speed, ViT-small takes 14.2 ms per frame, whereas ResNet50 takes 7.3 ms per frame, providing an important advantage for ResNet50.

In scale B, ViT-tiny has half the parameters of ResNet18 and runs twice as fast. However, compared to ViT-small, ViT-tiny had a performance drop of 2.4% in accuracy and 2.3% in F1-score. Similarly, there is a 2.2% difference in accuracy and a 2.1% difference in F1-score between ResNet50 and ResNet18. ViT-tiny is more efficient but sacrifices performance, while ResNet18 is slower and has more parameters than ResNet50.

When using ViT-based models, it is preferable to choose ViT small over ViT tiny if hardware and video frame rate allow it, due to the performance difference. Similarly, for CNN models, it is recommended to use ResNet50 instead of ResNet18, as the difference in inference speed and model size is less pronounced.

Our research stands out from prior studies (Yang et al., 2019; Yao et al., 2019; Xu et al., 2022) by developing and implementing a unique system on a real dairy farm. We have deployed a system which was built for commercial dairy farms, featuring a unique metal structure, accessible video cameras, a micro-processing edge device, and a router for convenient remote access.

Unlike the algorithms in previous studies that used manually collected data from handheld cameras, our system utilized data from fixed cameras positioned at a height of 4 m. This approach ensures alignment between the training data and the real-world implementation of the system.

Our system works at simultaneous detection and identification of multiple cow identities (up to 10). While Yao et al. (2019) also achieved this, the number of cows recognized concurrently remains unclear in their findings due to variable images. The proposed method offers another advantage with real-time video processing, ensuring zero latency when applying the algorithms through a dedicated low-cost hardware. By minimizing any disruption or interference, the system contributes to a favourable and ethical working environment for both the animals and the farmers.

We propose that future studies consider the adoption of an unsupervised clustering approach for cow facial identification, eliminating the need for pre-annotation and enabling transfer learning to scale up the system for implementation in larger dairy farms. Although various techniques were explored in this study, their full implementation was not realized, highlighting the significance of pursuing this path as a potential avenue for further research. Further endeavours can be directed towards enhancing the accuracy beyond our current best of 96.3%, using the same datasets. Simultaneously, there is room for improvement in reducing the inference speed for frames containing 10 faces by optimizing the DNN. It is still possible to achieve superior accuracy levels while concurrently accelerating the inference process for increased overall performance.

Conclusion

In this study, we developed a method that efficiently detects and recognizes multiple cow faces in real-time video files. It uses specialized DNNs, ensuring low latency, minimal memory usage, and high performance in under 20 ms per frame. For dairy farms, we recommend using YOLOv5-small for cow facial detection and ViT-small for facial recognition, providing excellent performance, memory efficiency, and fast inference speed. The implemented system does not have any physical contact with the animals, and it also does not interfere with the activities of the farmer. By minimizing the disruption or interference, the system can contribute to a more favourable and ethical working environment for both the animals and the farmers.

Ethics approval

All the procedures in this study were carried out in accordance with the accepted ethical and welfare standards of the Israel Ethics Committee (approval number IL-801/18).

Data and model availability statement

None of the data/models were deposited in an official repository. The data that support the study findings are confidential.

Declaration of Generative AI and AI-assisted technologies in the writing process

The authors did not use any artificial intelligence-assisted technologies in the writing process.

Author ORCIDs

Noam Bergman: https://orcid.org/0000-0003-2956-4646.
Yitzhak Yitzhaky: https://orcid.org/0000-0002-4974-9683.
Ilan Halachmi: https://orcid.org/0000-0002-2303-1016.

CRediT authorship contribution statement

N. Bergman: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Data curation, Conceptualization. **Y. Yitzhaky:** Writing – review & editing, Writing – original draft, Validation, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization. **I. Halachmi:** Writing – review & editing, Writing – original draft, Validation, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization.

Declaration of interest

The authors declare that there are no conflicts of interest.

Acknowledgements

Special thanks to Tom Lev-Ron for his advice and contribution on computer vision and deep-learning algorithms and engineering the system; to Joseph Lepar for substantial help in building the physical system and installing it; and to Assaf Godo for his help and advice in the design of the system.

Financial support statement

This study was supported by grants from the Israeli Chief Scientist of Agriculture, projects 459-451415 (“Kendel”), the Israel Dairy Board, fund number 459-4490 and the TechCare grant agreement number 459-6715Y51.

References

Alvarez, R.J., Arroqui, M., Mangudo, P., Toloza, J., Jatip, D., Rodríguez, J.M., Teyseyre, A., Sanz, C., Zunino, A., Machado, C., Mateos, C., 2018. Body condition estimation on cows from depth images using Convolutional Neural Networks. Computers and Electronics in Agriculture 155, 12–22. https://doi.org/10.1016/j.compag.2018.09.039.

Bezen, R., Edan, Y., Halachmi, I., 2020. Computer vision system for measuring individual cow feed intake using RGB-D camera and deep learning algorithms. Computers and Electronics in Agriculture 172,. https://doi.org/10.1016/j.compag.2020.105345 105345.

Bloch, V., Levit, H., Halachmi, I., 2019. Assessing the potential of photogrammetry to monitor feed intake of dairy cows. Journal of Dairy Research 86, 34–39. https://doi.org/10.1017/S0022029918000882.

Chen, Z., Xie, L., Niu, J., Liu, X., Wei, L. and Tian, Q., 2021. Visformer: The vision-friendly transformer. In: Proceedings of the IEEE/CVF international conference on computer vision, 11–17 October 2021, Montreal, QC, Canada, 589–598.

Deng, J., Dong, W., Socher, R., Li, L.-J., Kai L/, Fei-Fei, L., 2010. ImageNet: A large-scale hierarchical image database. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 20-25 June 2009, Miami, FL, USA, 4690–4699. doi: https://doi.org/10.1109/cvpr.2009.5206848.

Deng, J., Guo, J., Yang, J., Xue, N., Kotsia, I., Zafeiriou, S., 2022. ArcFace: Additive angular margin loss for deep face recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 44, 5962–5979.

Denholm, S.J., Brand, W., Mitchell, A.P., Wells, A.T., Krzyzelewski, T., Smith, S.L., Wall, E., Coffey, M.P., 2020. Predicting bovine tuberculosis status of dairy cows from mid-infrared spectral data of milk using deep learning. Journal of Dairy Science 103, 9355–9367. https://doi.org/10.3168/jds.2020-18328.

Devlin, J., Chang, M.-W., Lee, K., Toutanova, K., 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 3–5 June 2019, Minneapolis, MN, USA, 4171–4186. doi: https://doi.org/10.18653/v1/N19-1423.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N., 2020. An image is worth 16×16 words: Transformers for image recognition at scale. Retrieved on 19 November 2022 from https://arxiv.org/abs/2010.11929v2.

Geffen, O., Yitzhaky, Y., Barchilon, N., Druyan, S., Halachmi, I., 2020. A machine vision system to detect and count laying hens in battery cages. Animal 14, 2628–2634. https://doi.org/10.1017/S1751731120001676.

Grandini, M., Bagli, E., Visani, G., 2020. Metrics for Multi-Class Classification: an Overview. Retrieved on 02 December 2022 from https://arxiv.org/abs/2008.05756v1.

Halachmi, I., Guarino, M., Bewley, J., Pastell, M., 2019. Smart animal agriculture: application of real-time sensors to improve animal well-being and production. Annual Review of Animal Biosciences 7, 403–425. https://doi.org/10.1146/annurev-animal-020518-114851.

He, K., Zhang, X., Ren, S. and Sun, J., 2016. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, 26 June - 1 July 2016, Las Vegas, NV, USA, pp. 770–778.

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Adam, H., 2017. Mobilenets: Efficient convolutional neural networks for mobile vision applications. Retrieved on 11 January 2023 from https://arxiv.org/pdf/1704.04861.pdf.

Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q., 2017. Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, 22-25 July 2017, Honolulu, HI, USA, 4700–4708.

Jocher, G., 2020. YOLOv5. Retrieved on 22 October 2022 from https://github.com/ultralytics/yolov5.

Khan, S., Naseer, M., Hayat, M., Zamir, S.W., Khan, F.S., Shah, M., 2021. Transformers in vision: A survey. ACM Computing Surveys 54, 1–41. https://doi.org/10.1145/3505244.

Kumar, S., Tiwari, S., Singh, S.K., 2016. Face recognition of cattle: can it be done? Proceedings of the National Academy of Sciences, India Section A: Physical Sciences 86, 137–148. http://dx.doi.org/10.1007%2Fs40010-016-0264-2.

Lin T., 2015. LabelImg Git code. Retrieved on 31 January 2024 from https://github.com/tzutalin/labelImg.

Liu, W., Wen, Y., Yu, Z., Li, M., Raj, B., Song, L., 2017. Sphereface: Deep hypersphere embedding for face recognition. In; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 22–25 July 2017, Honolulu, HI, USA, 3738–6746. doi: https://doi.org/10.1109/CVPR.2017.713.

McGowan, J.E., Burke, C.R., Jago, J.G., 2007. Validation of a technology for objectively measuring behaviour in dairy cows and its application for oestrous detection. Proceedings of the New Zealand Society of Animal Production 67, 136–142.

Padilla, R., Passos, W.L., Dias, T.L.B., Netto, S.L., da Silva, E.A.B., 2021. A comparative analysis of object detection metrics with a companion open-source toolkit. Electronics 10, 279. https://doi.org/10.3390/electronics10030279.

Pastell, M., Frondelius, L., Järvinen, M., Backman, J., 2018. Filtering methods to improve the accuracy of indoor positioning data for dairy cows. Biosystems Engineering 169, 22–31.

Porto, S.M.C., Arcidiacono, C., Giummarra, A., Anguzza, U., Cascone, G., 2014. Localisation and identification performances of a real-time location system based on ultra-wide band technology for monitoring and tracking dairy cow behaviour in a semi-open free-stall barn. Computers and Electronics in Agriculture 108, 221–229.

Pu, H., Lian, J., Fan, M., 2018. Automatic recognition of flock behavior of chickens with convolutional neural network and kinect sensor. International Journal of Pattern Recognition and Artificial Intelligence 32, 1850023. https://doi.org/10.1142/S0218001418500234.

Qiao, Y., Su, D., Kong, H., Sukkarieh, S., Lomax, S., Clark, C., 2019. Individual cattle identification using a deep learning based framework. IFAC-PapersOnLine 52, 318–323. https://doi.org/10.1016/j.ifacol.2019.12.558.

Redmon, J., Divvala, S., Girshick, R., Farhadi, A., 2016. You only look once: Unified, real-time object detection. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 27-30 June 2016, Las Vegas, NV, USA, 779–788. https://doi.org/10.1109/CVPR.2016.91.

Ridnik, T., Ben-Baruch, E., Noy, A., Zelnik-Manor, L., 2021. ImageNet-21K Pretraining for the Masses. Retrieved on 05 November 2022 from https://doi.org/10.48550/arXiv.2104.10972.

Roessen, J., Harty, E., & Beirne, C. 2015. MooMonitor+ smart sensing technology and big data—Resting time as an indicator for welfare status on farms. In: Proceedings of ICAR Technical Series, 10 June 2015, Krakow, Poland, pp. 99–102.

Ros, G., Sellart, L., Materzynska, J., Vazquez, D., Lopez, A. M., 2016. The SYNTHIA Dataset: A Large Collection of Synthetic Images for Semantic Segmentation of Urban Scenes. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 27–30 June 2016, Las Vegas, NV, USA, pp. 3234–3243. https://doi.org/10.1109/CVPR.2016.352.

Shahriar, M.S., Smith, D., Rahman, A., Freeman, M., Hills, J., Rawnsley, R., Bishop-Hurley, G., 2016. Detecting heat events in dairy cows using accelerometers and unsupervised learning. Computers and Electronics in Agriculture 128, 20–26.

Shalloo, L., Dillon, P., Rath, M., Wallace, M., 2004. Description and validation of the moorepark dairy system model. Journal of Dairy Science 87, 1945–1959. https://doi.org/10.3168/jds.S0022-0302(04)73353-6.

Solawetz, J., 2021. YOLOv5 v6.0 is here – new Nano model at 1666 FPS. Retrieved on 12 January 2023 from https://blog.roboflow.com/yolov5-v6-0-is-here/.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z., 2016. Rethinking the Inception Architecture for Computer Vision. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 27–30 June 2016, Las Vegas, NV, USA, pp. 2818–2826. https://doi.org/10.1109/CVPR.2016.308.

Tan, M., Pang, R., Le, Q.V., 2020. EfficientDet: Scalable and efficient object detection. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 16–18 June 2020, Seattle, WA, USA, pp. 10778–10787. https://doi.org/10.1109/CVPR42600.2020.01079.

Tian, M., Guo, H., Chen, H., Wang, Q., Long, C., Ma, Y., 2019. Automated pig counting using deep learning. Computers and Electronics in Agriculture 163,. https://doi.org/10.1016/j.compag.2019.05.049 104840.

Tsai, H., Ambrogio, S., Narayanan, P., Shelby, R.M., Burr, G.W., 2018. Recent progress in analog memory-based accelerators for deep learning. Journal of Physics D: Applied Physics 51,. https://doi.org/10.1088/1361-6463/aac8a5 283001.

Wang, H., Wang, Y., Zhou, Z., Ji, X., Li, Z., Gong, D., Zhou, J., Liu, W., 2018b. Cosface: Large margin cosine loss for deep face recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 18–23 June 2018, Salt Lake City, UT, USA, pp. 5265–5274. https://doi.org/10.1109/CVPR.2018.00552.

Wang, H., Qin, J., Hou, Q., Gong, S., 2020. Cattle Face Recognition Method Based on Parameter Transfer and Deep Learning. In: Proceedings of 2nd International Conference on Computer Information, Science, and Artificial Intelligence (CISAI), 25–27 October 2019, Xi'an, China, 1453012054. https://doi.org/10.1088/1742-6596/1453/1/012054.

Wang, D., Tang, J.L., Zhu, W., Li, H., Xin, J., He, D., 2018a. Dairy goat detection based on Faster R-CNN from surveillance video. Computers and Electronics in Agriculture 154, 443–449. https://doi.org/10.1016/j.compag.2018.09.030.

Wen, Y., Zhang, K., Li, Z., Qiao, Y., 2016. A discriminative feature learning approach for deep face recognition. Vol. 9911 of Lecture Notes in Computer Science series, Springer-Verlag GmbH, Heidelberg, Germany. https://doi.org/10.1007/978-3-319-46478-7_31.

Werner, J., Leso, L., Umstatter, C., Schick, M., & O'Brien, B., 2017. Evaluation of precision technologies for measuring cows' grazing behaviour. In: Proceedings of the 19th Symposium of the European Grassland Federation, 7–10 May 2017, Alghero, Italy, pp. 82–84.

Wolfger, B., Timsit, E., Pajor, E.A., Cook, N., Barkema, H.W., Orsel, K., 2015. Accuracy of an ear tag-attached accelerometer to monitor rumination and feeding behavior in feedlot cattle. Journal of Animal Science 93, 3164–3168.

Wurtz, K., Camerlink, I., D'Eath, R.B., Fernández, A.P., Norton, T., Steibel, J., Siegford, J., 2019. Recording behaviour of indoor-housed farm animals automatically using machine vision technology: A systematic review. PloS One 14, e0226669.

Xu, B., Wang, W., Falzon, G., Kwan, P., Guo, L., Sun, Z., Li, C., 2020. Livestock classification and counting in quadcopter aerial images using Mask R-CNN. International Journal of Remote Sensing 41, 8121–8142. https://doi.org/10.1080/01431161.2020.1734245.

Xu, B., Wang, W., Guo, L., Chen, G., Li, Y., Cao, Z., Wu, S., 2022. CattleFaceNet: a cattle face identification approach based on RetinaFace and ArcFace loss. Computers and Electronics in Agriculture 193,. https://doi.org/10.1016/j.compag.2021.106675 106675.

Yang, Z., Xiong, H., Chen, X., Liu, H., Kuang, Y., Gao, Y., 2019. Dairy Cow Tiny Face Recognition Based on Convolutional Neural Networks. Vol. 11818 in Lecture Notes in Computer Science series, Springer-Verlag GmbH, Heidelberg, Germany. https://doi.org/10.1007/978-3-030-31456-9_24.

Yang, Q., Xiao, D., Lin, S., 2018. Feeding behavior recognition for group-housed pigs with the Faster R-CNN. Computers and Electronics in Agriculture 155, 453–460. https://doi.org/10.1016/j.compag.2018.11.002.

Yao, L., Hu, Z., Liu, C., Liu, H., Kuang, Y., Gao, Y., 2019. Cow face detection and recognition based on automatic feature extraction algorithm. In: Proceedings of the ACM Turing Celebration Conference, 17–19 May 2019, Chengdu, China, pp. 1–5. https://doi.org/10.1145/3321408.3322628.

Zaidi, S.S.A., Ansari, M.S., Aslam, A., Kanwal, N., Asghar, M., Lee, B., 2022. A survey of modern deep learning-based object detection models. Digital Signal Processing 126,. https://doi.org/10.1016/j.dsp.2022.103514 103514.

Zhang, X., Fang, Z., Wen, Y., Li, Z., Qiao, Y., 2017. Range Loss for Deep Face Recognition with Long-Tailed Training Data. In: Proceedings of the IEEE International Conference on Computer Vision, 22–29 October 2017, Venice, Italy, 5419-5428. https://doi.org/10.1109/ICCV.2017.578.

Zhang, Q., Yang, L.T., Chen, Z., Li, P., 2018. A survey on deep learning for big data. In Information Fusion 42, 146–157. https://doi.org/10.1016/j.inffus.2017.10.006.

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