A computer vision approach based on deep learning for the detection of dairy cows in 2 free stall barn

Published Version:
Tassinari, P., Bovo, M., Benni, S., Franzoni, S., Poggi, M., Mammi, L.M.E., et al. (2021). A computer vision approach based on deep learning for the detection of dairy cows in 2 free stall barn. COMPUTERS AND ELECTRONICS IN AGRICULTURE, 182, 1-15 [10.1016/j.compag.2021.106030].

Availability:
This version is available at: https://hdl.handle.net/11585/800275 since: 2021-02-16

Published:
DOI: http://doi.org/10.1016/j.compag.2021.106030

(Article begins on next page)

A computer vision approach based on deep learning for the detection of dairy cows in free stall barn

Patrizia Tassinari[a], Marco Bovo[a] *, Stefano Benni[a], Simone Franzoni[b], Matteo Poggi[b], Ludovica Maria Eugenia Mammi[c], Stefano Mattoccia[b], Luigi Di Stefano[b], Filippo Bonora[a], Alberto Barbaresi[a], Enrica Santolini[a], Daniele Torreggiani[a]

[a] Department of Agricultural and Food Sciences, University of Bologna, Bologna, Italy
[b] Department of Computer Science and Engineering, University of Bologna, Bologna, Italy
[c] Department of Veterinary Medical Sciences, University of Bologna, Bologna, Italy
* **Corresponding author**. Email: marco.bovo@unibo.it

Abstract

Precision Livestock Farming relies on several technological approaches to acquire in the most efficient way precise and up-to-date data concerning individual animals. In dairy farming, particular attention is paid to the automatic cow detection and tracking, as such information is closely related to animal welfare and thus to possible health issues. Computer vision represents a suitable and promising method for this purpose.

This paper describes the first step for the development of a computer vision system, based on deep learning, aiming to recognize in real-time the individual cows, detect their positions, actions and movements and record the time history outputs for each animal.

Specifically, a neural network based on deep learning techniques has been trained and validated on a case study farm, for the automatic recognition of individual cows in videos recorded in the barn. Four cows were selected to train and validate a YOLO neural network able to recognize a cow starting from the coat pattern. Then, precision-recall curves of the identification of individual cows were elaborated for both the specific target classes and the whole dataset in order to assess the performances of the network.

By means of data augmentation techniques, an enlarged dataset has been created and considered in order to improve the performance of the network and to provide indications to increase detection efficiency in those cases where data acquisition is not easy to be carried out for long periods. The mean average precision of the detection, ranging from 0.64 to 0.66, showed that it is possible to properly identify individual cows based on their morphological appearance and that the piebald spotting pattern of a cow's coat represents a clearly distinguishable object for a computer vision network. The results also led to obtain indications about the quantity and the characteristics of the images to be used for the network training in order to achieve efficient detections when facing with applications involving animals.

Keywords: precision livestock farming; computer vision; deep learning; dairy cow; herd management

List of symbols

Symbol	Description
A_t	Average Area of the bounding boxes in training phase
A_v	Average Area of the bounding boxes in validation phase
O_t	number of Occurrences in training phase
O_v	number of Occurrences in validation phase
tp	true positive
tn	true negative
fp	false positive
fn	false negative
P	Precision
R	Recall
F_1	F_1-score
Pr	detection Probability
C	Confidence score
AP	Average Precision
mAP	mean Average Precision
IoU	Intersection over Union
AIoU	Average Intersection over Union

1. Introduction

The implementation of precision livestock farming (PLF) techniques in animal husbandry involves many fields of the technological innovation and several researchers are currently seeking to apply new methodologies and algorithms for both commercial and research purposes (Tullo et al., 2019). In particular, with reference to innovative applications in the livestock sector, the animals are increasingly being analyzed and studied with the help of informatics tools such as support vector machines, random forests techniques, neural networks, machine learning approaches etc. (Kamilaris and Prenafeta-Boldú, 2018; Tsai and Huang, 2014; Li et al., 2017; Okura et al., 2019). In this context, multiple challenges involve the dairy cattle sector, where the methods currently available are often unsuitable to manage the collected data and to fully extrapolate the potential informative content.

In fact, weather stations monitoring barn temperature and humidity, robotic milking systems helping the daily work of the farmers, collars and pedometers controlling animals' activities and positions (Berckmans, 2014) are capable to collect huge amounts of real-time data currently used to support the herd management.

PLF has thus contributed to switch the analysis framework of a farm from a data-poor to a data-rich situation: the research key challenge is actually to turn those data into knowledge able to provide decision-maker with real-time support for the cattle farm optimization (Barkema et al., 2015; Bewley et al., 2017; Fournel et al., 2017; Van Hertem et al., 2014; Halachmi et al., 2013; Guzhva et al., 2016; Martinez-Ortiz et al., 2013).

Several researches have pointed out the opportunity to develop both algorithms suitable to provide early warning and control systems able to optimize animal welfare and productivity based on data collected through Information Communication Technology (ICT) systems (Alsaaod et al., 2019; Jaeger et al., 2019; Cowley et al., 2015). In this context, computer vision together with numerical analysis methodologies proved to have fundamental importance (Van Hertem et al., 2018) . Computer Vision techniques are meant to be applied in this research field in order to automate actions normally carried out by the human visual system (Taigman et al., 2014). The aim of these algorithms is to "teach" a computer to apprehend from images and videos in order to simulate the human vision and substitute the human beings in repetitive or complex actions. They have already been tested on animals with promising results (Norouzzadeh et al., 2017; Trnovszky et al., 2017), also in the livestock farming area (Aydin, 2017; Van Hertem et al., 2013), but recognizing each individual cow within the herd is still representing a challenging issue.

The monitoring of position and movements of the individual animals may be necessary to quantify the main indices related to animal welfare (Song et al., 2008; Jiang et al., 2019) and behavior (Porto et al., 2013, 2015), as well as to identify any preferences of the cows regarding different zones of the barn.

Moreover, video monitoring of the herd, together with the adoption of quantitative criteria to control animal welfare by means of computer vision, may represent a tool to improve citizens' consciousness about rearing conditions and increase their knowledge about feeding and housing practices. A recent study found that fresh food and water, pasture access, gentle handling, space, shelter, hygiene, fresh air and sunshine, social companions, absence of stress, health and safety from predators are considered by citizens as necessary requirement for dairy cattle "good life" (Ventura et al., 2016). These results suggested that a transparent exposure of livestock farming to the public may resolve some concerns, and video recording appear to be a powerful tool for an effective and widespread information.

Therefore, the monitoring requirements can be considered according to two main levels of information and complexity. The first one concerns the identification at regular time intervals of the number of animals that are in a certain position, for example lying in a cubicle, standing at the manger etc. The second one deals with the identification, instant by instant, of the behavior of the individual cows, with the possibility of calculating, for each animal, the time spent in each position and the temporal sequence of its positions, including the trajectory of its movements.

The first level of information makes it possible to quantify the aforementioned indices and to have an overview of the performances of the general animal welfare conditions in the farm, by appropriately integrating these indices with information on feeding and productivity of the animals, which can be deduced from other sources, such as the milking robots and mixed ration delivery systems. The second level of information allows to have specific information on the individual cow welfare condition and on the use of the various areas of the barn and represents a challenge for innovation in PLF. A computer vision system implemented through deep learning constitutes a suitable approach to achieve the latter objective.

The aim of this study is to develop and test the reliability level of a computer vision system, based on deep learning techniques, for the automatic recognition of individual cows within images representing their position. In particular, whilst developing a new software framework lies outside the scope of this paper, the paper focuses on methodological aspects related to a more efficient application of ICT in the dairy cows monitoring field. In fact, while object detection is already applied in commercial applications in various contexts, individual cow recognition still represents an open issue for both the research and commercial fields, since no consolidated approach still exists. Actually, some of these blocks of information are collected by means of different very expensive sensors (ALLFLEX, 2020; DELAVAL, 2020; AFIMILK, 2020), the most of them to be worn by the animals but with the system proposed here, it will be possible to collect wider information, with a less expensive technology and finally yet importantly using a system that avoids the problems and labor due to wearable sensors.

Therefore, the paper focuses on the selection, validation and performance assessment of a neural network, in terms of speed and accuracy, and at the same time outlines the key issues of the broader process, which is strictly depending and related to the enabling steps presented in the paper. Further testing and validating cow detection procedures in various contexts, in different types of livestock structures, and in different operating conditions, is an important research field, contributing to the definition of consolidated approaches enabling cow recognition, displacement tracking and cow action/behavior (eating, drinking, lying down, standing etc.) recognition systems. This system could represent an innovative and useful tool to support the farmer in the daily management and decision-making. In fact, by means of the technology proposed here, it will be possible to calculate, for example, the indices connected to animal welfare but also to check the time spent for nutrition, drinking and walking. The outcomes of this monitoring could be effectively used by the farmers to identify potential anomalies or diseases and promptly apply specific corrective actions (e.g. on the fans controlling the barn environmental conditions, water supply, etc.). In addition to the technological system and the experimental setup adopted, the paper describes an innovative algorithm for the detection of the cows implemented and tested on a case study farm with computer vision procedures. The promising results reported here represents a first preliminary contribution to the progress of the computer vision for herd monitoring applications and could be an important support for the following study of the movements of the cows in the barn and for the analysis of their actions and behavior.

2. Materials and methods

2.1 Study case

The case study considered in this work is the experimental dairy cattle farm of the University of Bologna, located in Ozzano Emilia, Bologna, in the North-East of Italy, where Holstein Friesian cattle is reared. This farm is managed by the Department of Veterinary Medical Science and represents a unique reality in the national context, thanks to its equipment and monitoring systems, which will allow to carry out the integrated analyses aimed at the definition of models for the interpretation of milk production, reproductive, environmental and management data. The building is a free stall experimental barn hosting about 150 animals, including 80 cows (72 milking and 8 dry cows, and 70 among calves and heifers). The barn has a resting area in bedding material for dry cows and litter cubicles for lactating cows (see Figure 1). A layout of the building is showed in Figure 1.

The barn is provided with cooling ventilation systems based on the Temperature-Humidity Index (THI) value. The cows are milked twice a day and for each cow, the behavioral, productive and health parameters are daily recorded automatically. The animals are fed with total mixed ration and auto-feeders for concentrate supplementation.

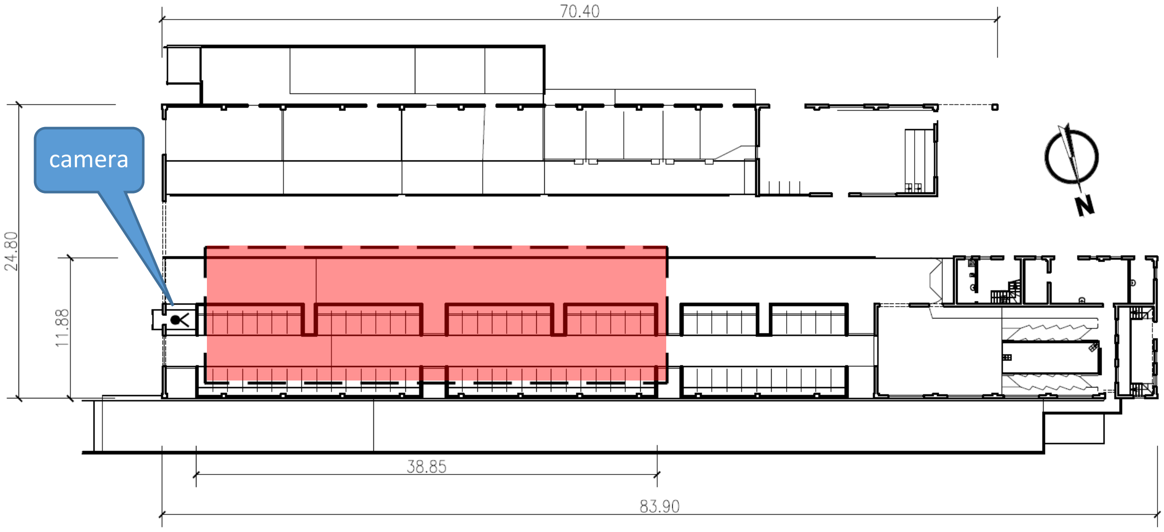


Figure 1. Layout of the experimental barn. The dashed line delimitates the area framed by the camera (red colored).

2.2 *Software and neural network*

2.2.1 *Tagging software*

VoTT (Visual object Tagging Tool) (Microsoft, 2018) was the software selected for the tagging tasks. It is an open source annotation and labelling tool for image and video assets, it is a React + Redux Web application written in Typescript. This software has been selected because it has several features, useful for the application in the field of the present work. For example, it has an extensible model for importing data from local or cloud storage providers and an extensible model for exporting labelled data to local or cloud storage providers. Moreover, it has been developed not only to analyze images, but also video frames. So, it is possible to decide the interval of frames per second and VoTT automatically transforms a video in a set of pictures. VoTT was programmed following the 'Bring Your Own Data' (BYOD) approach and in VoTT, connections are used to configure and manage source, i.e. the assets to label, and the target, i.e. the location to which labels should be exported. Then, for the development of the computer vision technology in the field of animal monitoring, VoTT results very practical to set up and facilitate an end-to-end machine learning pipeline. In this work the version 1.0.8 has been used.

2.2.2 *Detection algorithm and deep learning network*

At the state-of-the-art, in the field of computer vision for object detection, different more or less efficient algorithms could be considered. For example, region-based convolutional neural networks (R-CNNs) (Girshick et al., 2014) have been a pioneering approach that applies deep models to object detection. Actually, Faster R-CNN (Ren et al., 2017) has been proved to be one of the most performing of its class, if compared with R-CNN and Fast R-CNN (Girshick, 2015) and it has been already applied for cow detection for instance by (Bezen et al., 2020). Different is the approach implemented in YOLO - You Only Look Once (Redmon et al., 2016), where predictions are made with a single network evaluation, thus making the process much faster than region-based convolutional neural networks (R-CNN) which require thousands of evaluations for a single image. YOLO presents a totally different approach from prior detection systems, which proposes classifiers or localizers to perform detection. It is one of the state-of-the-art detectors which are capable of localizing and classifying multiple objects in images. In particular, the detector is faster than two-stage detectors: while they propose object regions first and investigate the regions for object localization and classification, YOLO combines the two stages into one neural network. Therefore, instead of applying the model to an image at multiple locations and scales, so that high-score regions of the image are considered detections, Yolo applies a single neural network to the full image. This network divides the image into regions and predicts bounding boxes and probabilities for each region and these bounding boxes are weighted by the predicted probabilities. The improvements in YOLO v3 (Redmon and Farhadi, 2018) made the algorithm even faster and suitable for problems like those investigated here. In fact, the future real applications of this system will face with the detection of large number of cows in a herd, during the time, and should be able to recognize the possible action the cow is doing in real-time. For this reason a fast algorithm is necessary and this has driven the choice of the authors towards the framework Darknet (Redmon, 2013), an open source framework for neural network development including YOLO v3. The algorithm has the advantage to look at the whole image at test time, so that predictions are informed by the global context of the image (see Figure 2). More specifically, YOLO v3 predicts an objectness score for each bounding box using logistic regression. This score should be 1 if the bounding box prior overlaps a ground truth object by more than any other bounding box prior. Each box predicts the classes the bounding box may contain using multi-label classification. At the current state of the research the authors selected and adopted YOLO v3 as detection system but the evaluation of the performances of different detection algorithms, for the problem investigated here, will be object of future investigations.

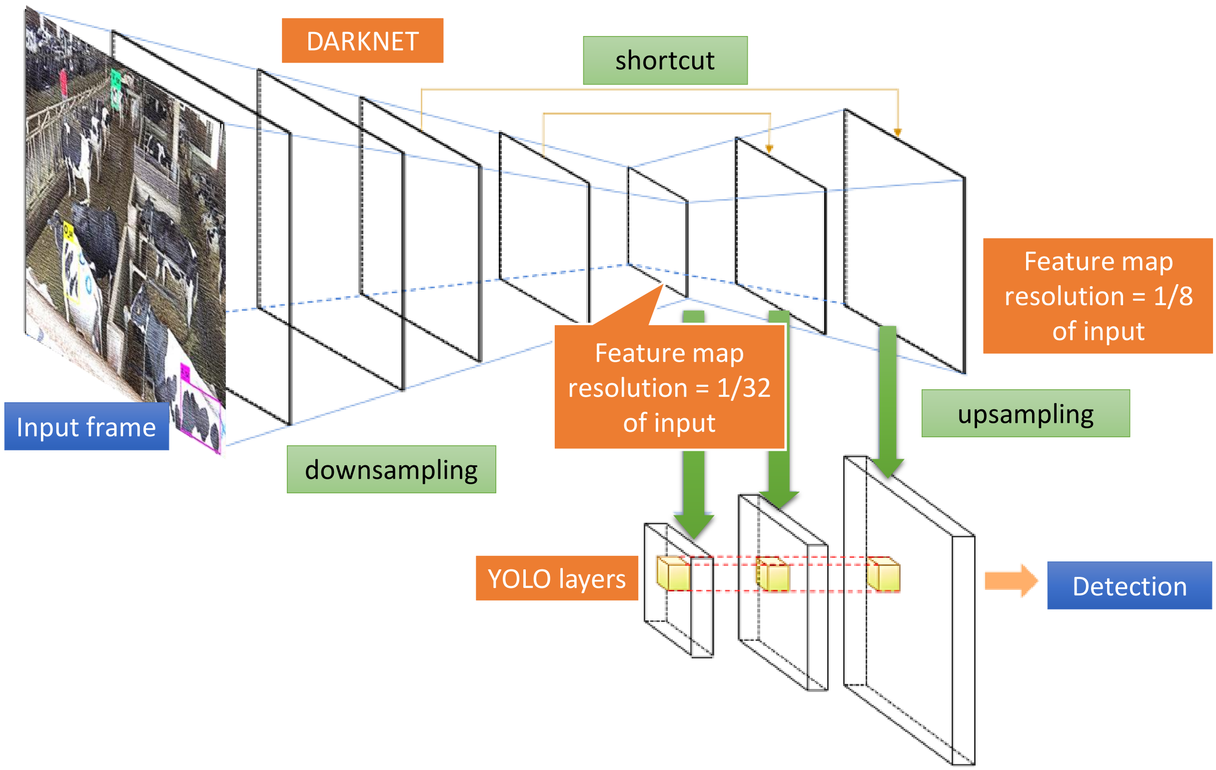


Figure 2. Architecture of YOLO v3.

2.3 *Data collection*

As the performance of the neural network are strictly connected to the quantity and the quality of the training dataset, particular attention was paid to the collection of a suitable dataset of video frames. The videos were registered by a HDR-CX115E (Sony) camera (see Figure 3) in a high quality standard (HD resolution, 25 frames per second). The recording has been conducted on a tripod positioned 2 meters above the barn floor, so the total height for the recording was about 3.50 meters.

The section recorded by the camera focused on the feeding area, including the rack (on the left) and the cubicles (on the center and on the right of the frames). A limitation of this position is that the images of the cows up in the trough are greater for quality and quantity, so the dataset is composed by more photos of the left hips than the right ones.

Figure 3. The HDR-CX115E camera positioned for video recording.

2.4 Research method

The research has been structured in the following main phases:

1) random selection in the herd of a sample of 10% of the cows' population in the free stall area monitored by the camera (i.e. about 40 cows) during the recording phase. The sample is suitable to verify if the network is capable to recognize a cow among others and to distinguish between more animals. Then the animals have been marked, with a specific blue paint, with a letter only to help the identification of the cows in the different frames. It is worth to notice this aspect since the neural network has not been

trained based on these symbols but the bounding boxes considered only pelt portions with natural pattern of the cows;

2) recording of the videos after selecting the more suitable position for the camera;
3) creation of the dataset for the training phase;
4) training of the neural network, i.e. definition of the parametric weights for object recognition;
5) validation test of the neural network with the weights defined in phase 4 and scoring the results by means of the performance indicators defined before;
6) creation of an augmented "virtual" dataset aiming to improve the detection performances of the network for the classes poorly represented in the frames used for the network training;
7) repetition of the phases from 3 to 5 in order to assess the improvement in the detection performances after the manipulations operated to the frame dataset.

2.5 Experimental setup and tests

Four cows have been selected to train and test the neural network adopted in the study. The four letters X, V O and I have been used only to identify a specific cow and the letter corresponding to each cow was drawn on both right and left hips and on the forehead of the various cows. Two types of classes have been defined for the identification of each cow corresponding to the pelt of the hips of the four cows. A total number of 8 classes (4 cows × 2 hips) has been adopted for the neural network training/validation in order to recognize the cows by the black-white pattern of each specific pelt. Therefore, each class was identified by the capital letter indicating the cow followed by right (r) or left (l) for identify the two different hips. The eight classes considered in the study are: X_{left}, X_{right}, V_{left}, V_{right}, O_{left}, O_{right}, I_{left} and I_{right}. For example, Figure 4 illustrates the internal view of the barn with a detail of the blue paint on the right hip of the cow labelled "V", thus labelled as class V_{right}.

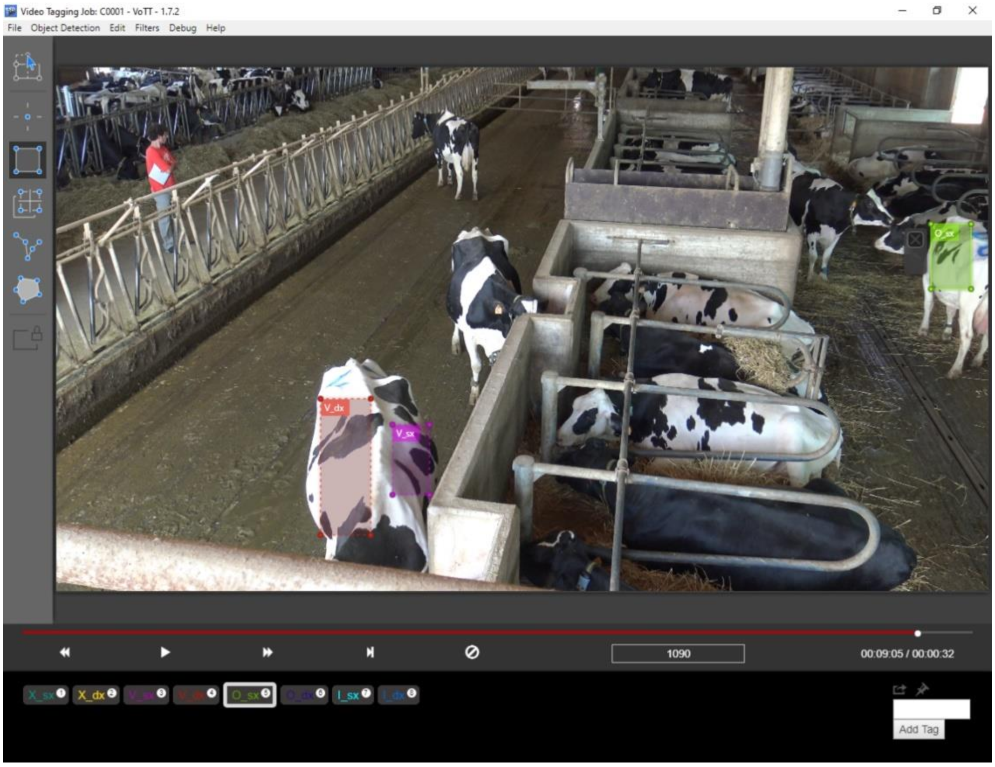
Figure 4. Example of a frame with detail of the blue paint on the right hip of the cow labelled "V" so representing the class V_{right}.

The videos were recorded on July 2019, collecting a total duration of 210 min to carry out the study. The frame tagging phase was performed through the abovementioned VoTT software, which allowed to sample the videos at a chosen frequency and to make tags (with bounding boxes) on all the frames. In this study, the sampling frequency for the selection of the frame dataset was chosen equal to 2 frames per second because the scenario does not change in a fast way and so a higher frequency would have been redundant. Therefore, about 25200 frames were sampled. The bounding boxes used for tagging the frames were rectangular, rather than square, because the objects to be tagged, i.e. the pelt of the cows, had generally different horizontal and vertical dimensions. The coat area selected in the tagging bounding box has been the biggest rectangular area (with horizontal orientation) identifiable with continuity within the image of the hip of the cow.

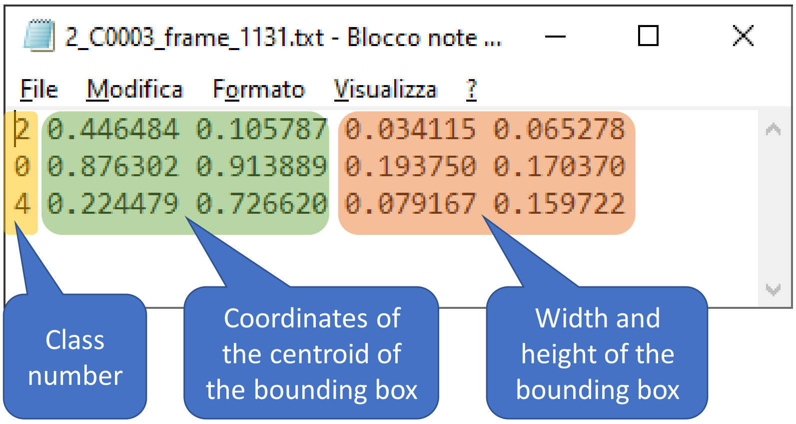
Thus, at the end of the tagging phase we obtained:

- a collection of graphical files corresponding to every sampled frame
- for each frame, a text data file like the one in Figure 5 containing the class number recognized in the frame (if any), the coordinates of the centroid of the corresponding box and the sizes of the box.

A total of 11754 frames showing at least one of the classes were identified and labelled.



(a)



(b)

Figure 5. Tagging phase through VoTT. (a) Example of graphical file of a frame and (b) example of the data acquired by the tagging process. The coordinates of the centroid and the dimensions of the box are expressed as ratios of the dimensions of the frame.

The data acquired through the tagging phase were then analyzed to quantify the occurrences of the various target classes. Moreover, the sizes (i.e. width and height) of the bounding boxes were computed as an indicator of the visibility of the target cow within the frame. In fact, a small box may indicate either a position of the cow far from the recording viewpoint, or a partial coverage of the animal by an object in the foreground.

The data about the occurrences of the target classes were considered for a proper definition of the split of the dataset in a training set and a validation set. The criterion adopted for the split was the selection of sets of consecutives frames accounting for about 80% of the occurrences for each class in the training set and the remaining 20% for the validation set. 10105 frames for training, 1649 frames for validation. The data resulting for each class are summarized in Table 1.

Table 1
Data resulting from the analysis of the tagging process.

Class code	Class #	Occurrences in the whole dataset
X_{left}	0	1649
X_{right}	1	1575
V_{left}	2	3249
V_{right}	3	839
O_{left}	4	1113
O_{right}	5	649
I_{left}	6	2524
I_{right}	7	771

The training and validation of the network was performed by a Nvidia GTX GeoForce 12GB Titan GPU. The already trained weights downloaded from the Darknet repository (Redmon and Farhadi, 2018) were used as initial weights for training and 10 000 iterations were performed to obtain the final (adjusted) weights.

2.6 Results assessment

In this subsection the metrics adopted for the evaluation of the performance of the system are presented. The first metric to introduce is the intersection-over-union (IoU) index, also known as Jaccard index (Rezatofighi et al., 2019), maybe the most commonly used metric for comparing the similarity between two general images. IoU encodes the properties of the items under comparison (e.g. widths, heights, locations of bounding boxes) and then calculates a normalized measure reported in Eq. (1) as the ratio between the area of the intersection divided by the union of the two bounding boxes (i.e. the predicted and the ground truth bounding boxes).

$$\mathrm{IoU} = \frac{\text{Area of ground truth box} \cap \text{Area of predicted box}}{\text{Area of ground truth box} \cup \text{Area of predicted box}} \tag{1}$$

IoU results invariant to the problem scale and thanks to this feature the most of the performance measures in segmentation, object detection, and tracking are based on this metric (Rezatofighi et al., 2019).

In pattern recognition applications, the precision (P) is the fraction of relevant instances among the retrieved instances, while the recall (R) is the fraction of relevant instances that have been retrieved over the total amount of relevant instances. Both precision and recall are therefore based on an understanding and measure of the relevance. Precision (*P*) and recall (*R*) can be expressed, in analytical form, by means of the following expressions:

$$P = \frac{tp}{tp+fp} \tag{2}$$

$$R = \frac{tp}{tp+fn} \tag{3}$$

Where: tp represents true positive, i.e., the number of cases that the detector successfully detects a class in an image with IoU greater than a prescribed threshold; fp is false positive, i.e., the number of cases that the detector reports other objects as a target class in an image, or IoU is less than a prescribed threshold; fn is false negative, i.e. the number of cases that the detector fails to detect a target class in an image. In the present work, the specific threshold has been fixed equal to 0.5. Precision is also known as “positive predictive value”, while recall is also called “true positive rate” or “sensitivity” and this last represents the proportion of actual positives are correctly identified.

In the interpretation of computer vision results, the balanced F_1-score (also F-score or F-measure) is a metric that combines both P and R and represent the harmonic mean (Nie et al., 2019):

$$\mathrm{F}_1 = 2 \cdot \frac{\mathrm{P} \cdot \mathrm{R}}{\mathrm{P}+\mathrm{R}} = \left(\frac{\mathrm{P}^{-1}+\mathrm{R}^{-1}}{2}\right)^{-1} \tag{4}$$

This metric coincides with the square of the geometric mean divided by the arithmetic mean of precision and recall and is clearly close to the arithmetic mean of the two when P and R have similar values. F_1 reaches its best value at 1.0 and the worst at 0.0.

Moreover, the confidence score C (%) has been considered, which quantifies the reliability of the recognition of a given object within a frame. Confidence score can be calculated using the formula:

$$C = Pr \times IoU \tag{5}$$

where: Pr represents the detection probability assessed by the network that the object at hand belongs to the class attributed to it.

Precision and recall have been computed for each class based on different confidence thresholds ranging from 0 to 100%, and the precision-recall curves have been drawn for each class. Besides, AP (average precision) is a popular metric measuring the accuracy of object detectors. AP computes the average precision value for recall value from 0 to 1 for a specific class (Szeliski, 2011). Therefore, AP can be computed as the area under the P-R curve of such class. Then, the mean average precision (mAP) of the network was assessed as the mean of the AP values of the different classes. With analogous criteria is possible to define the average IoU (AIoU) for a specific class as the average of all the IoU values of the occurrences of the same class.

3. Results and Discussion

3.1 *With original data frames*

This subsection deals with the main results obtained by considering the original set of frames described in the previous section. The set is constituted by 11754 frames, 10105 (about 85% of the total dataset) have been used for the network training whereas 1649 (about 15%) for the network validation test. The frames to be used in the validation phase were carefully selected in order to guarantee that all the 8 considered classes were adequately represented in the frames. The occurrences of each class are reported in Table 2, for both training and validation phases. Obviously, the sum of the occurrences of all the classes, i.e. 10167 and 2202 respectively for training and validation phases, are bigger than the total frame number since some frames includes multiple cows belonging to different classes.

Table 2
Number of occurrences and average area of the bounding box for each class and for both training and validation original datasets.

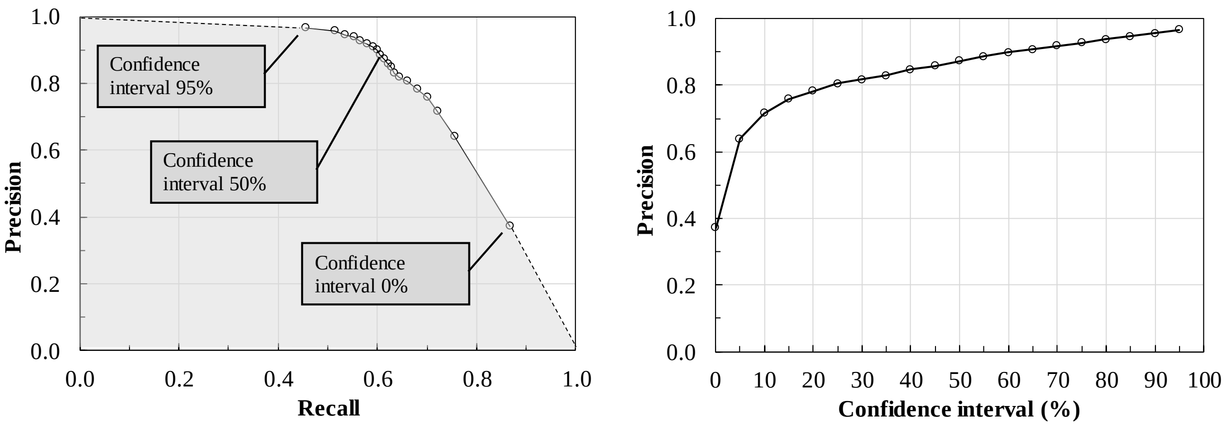
Training dataset									
	X_{left}	X_{right}	V_{left}	V_{right}	O_{left}	O_{right}	I_{left}	I_{right}	Sum
O_t	1325	1273	2599	673	904	512	2270	611	10167
A_t	0.01063	0.00920	0.00540	0.01688	0.00538	0.01567	0.00447	0.00502	-
$O_t \times A_t$	14.08	11.71	14.03	11.36	4.86	8.02	10.15	3.07	-
Validation dataset									

	X_{left}	X_{right}	V_{left}	V_{right}	O_{left}	O_{right}	I_{left}	I_{right}	Sum
O_v	324	302	650	166	209	137	254	160	2202
A_v	0.02748	0.01540	0.00356	0.01306	0.01542	0.01364	0.01552	0.01275	-
$O_v \times A_v$	8.90	4.65	2.31	2.17	3.22	1.87	3.94	2.04	-

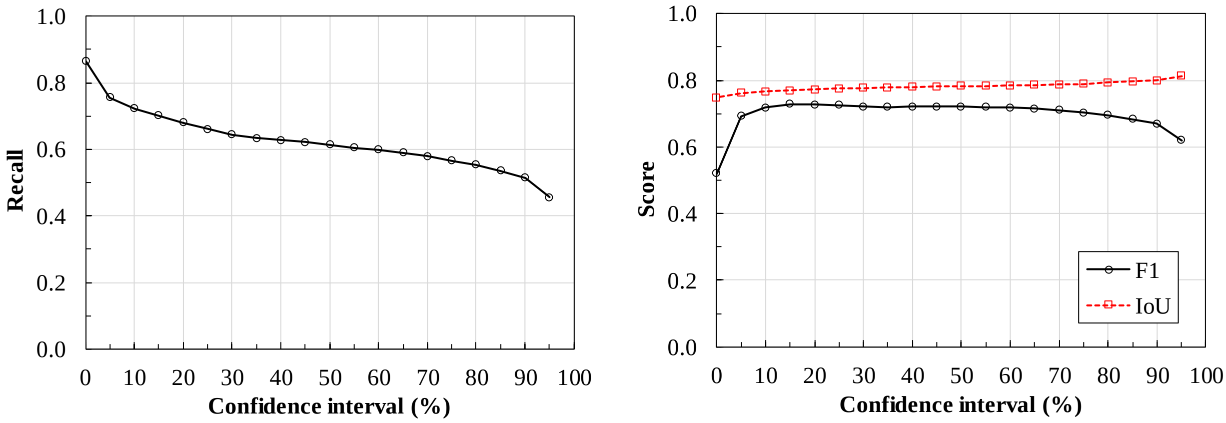
The validation of the computer vision detection could be carry out from a visual (or graphical) point of view, by looking if in one specific frame the classes object of the test (in this case the 8 hips of the 4 cows) are properly identified by the neural network. For example, Figure 6 shows in the yellow box, in the magenta box and in the green box, the identification of the left hip respectively of the cow O, cow X and cow V. The accuracy of the detection, reported in the rectangle at top-right of the figure, represent the confidence score C for each class, as calculated by YOLO, and could be correlated to the probability of finding the cow in the bounding box. For the frame in the Figure 6, for example, the detection results very good since the class O_{left} and V_{left} have C of 100% whereas X_{left} has C about 98%.

Figure 6. Example of visual validation of the classes in a frame. The yellow box (O_sx) is the identified left hip of the cow marked with the letter O; the magenta box (X_sx) is the identified left hip of the cow X and the green box (V_sx) is the identified left hip of cow V.

In addition, the validation of the neural network can be performed from a mathematical point of view on the basis of global P and R scores (i.e. considering the whole data set coming from the various classes). As an example, Figure 7(a) shows the global P-R graph (i.e. considering the dataset of the all 8 classes considered in the study) of the validation test. The trend of the global P-R graph was obtained by considering 20 different confidence interval (CI) with increasing confidence level, from 0.0% to 95% with step 5%, for the assessment of the detections based on the IoU of the single detection. As a general trend, the lower confidence interval produces points with low P and high R values. The opposite for high confidence interval. It seems useful to remember that the optimal graph trend should be that presenting high level of precision (i.e. higher than 0.8) all along the R value. Figures 7(b) and 7(c) respectively show the trend of P and R for the different CI values. For the case at hand the P value is adequate (i.e. higher than 0.8) for CI higher than 20%. For CI equal to 20% the R value is about 0.7 and it means the neural network is able to detect about the 70% of the “real” occurrences of the different classes for the various frames. Table 3 reports the main data resulting from the validation phase and related to the whole dataset. Moreover, same table reports the global F_1-score and IoU values for each CI, also depicted in Figure 7d. IoU values go from 0.75 to 0.81 with an AIOU equal to 0.78. In the object detection field values higher than 0.7 until 1.0, commonly, identified detections good to excellent, then we reached, on average, a rather good detection from the network.



(a) (b)



(c) (d)

Figure 7. Principal trends obtained from the validation test by considering all the occurrences dataset and reported for different confidence interval (with original data frames). (a) P-R curve. (b) P trend; (c) R trend; (d) F_1 and IoU score for different confidence interval.

Table 3

Main results from the validation test by considering all the occurrences dataset and reported for different confidence interval (with original data frames).

Confidence interval	True positive*	Ground truth**	Precision	Recall	F_1-score	IoU
0.95	1004	1039	0.9663	0.4559	0.6196	0.8126
0.90	1134	1186	0.9562	0.5150	0.6694	0.8000
0.85	1178	1245	0.9462	0.5350	0.6835	0.7960
0.80	1218	1299	0.9376	0.5531	0.6958	0.7928
0.75	1247	1346	0.9264	0.5658	0.7026	0.7891
0.70	1277	1390	0.9187	0.5795	0.7107	0.7874
0.65	1301	1434	0.9073	0.5904	0.7153	0.7853
0.60	1319	1467	0.8991	0.5985	0.7187	0.7838
0.55	1333	1505	0.8857	0.6049	0.7189	0.7833
0.50	1353	1551	0.8723	0.6140	0.7207	0.7821
0.45	1370	1596	0.8584	0.6217	0.7211	0.7811
0.40	1382	1631	0.8473	0.6272	0.7208	0.7803
0.35	1398	1686	0.8292	0.6344	0.7188	0.7790
0.30	1422	1740	0.8172	0.6449	0.7209	0.7774
0.25	1458	1812	0.8046	0.6612	0.7259	0.7752
0.20	1500	1918	0.7821	0.6803	0.7276	0.7723
0.15	1548	2040	0.7588	0.7012	0.7289	0.7700
0.10	1598	2234	0.7153	0.7221	0.7187	0.7665
0.05	1681	2632	0.6387	0.7566	0.6926	0.7617
0.00	2011	5409	0.3718	0.8665	0.5203	0.7480

* : is the number of true positive occurrences detected from the neural network

** : is the number of “real” occurrences in the dataset as resulting from the visual detection performed by the operator.

As far as the single class is concerned, Figure 8 shows the Precision-Recall graphs for every considered class and in Table 4 the most important parameters are collected, useful to judge the detection quality of each single class. In fact, if in some contexts an “on-average” detection score is sufficient (Szeliski, 2011). In the present applications, it is not enough being the single class detection score, important as much the “on-average” score for practical PLF purposes. Then, it is possible to identify in Figure 8, and evaluate from Table 4, the classes with better/worse detection scores. E.g., from the table, the classes with better AP are V_{left}, X_{right} and O_{left}. Instead, the classes with the worst AP are O_{right}, I_{right} and I_{left}.

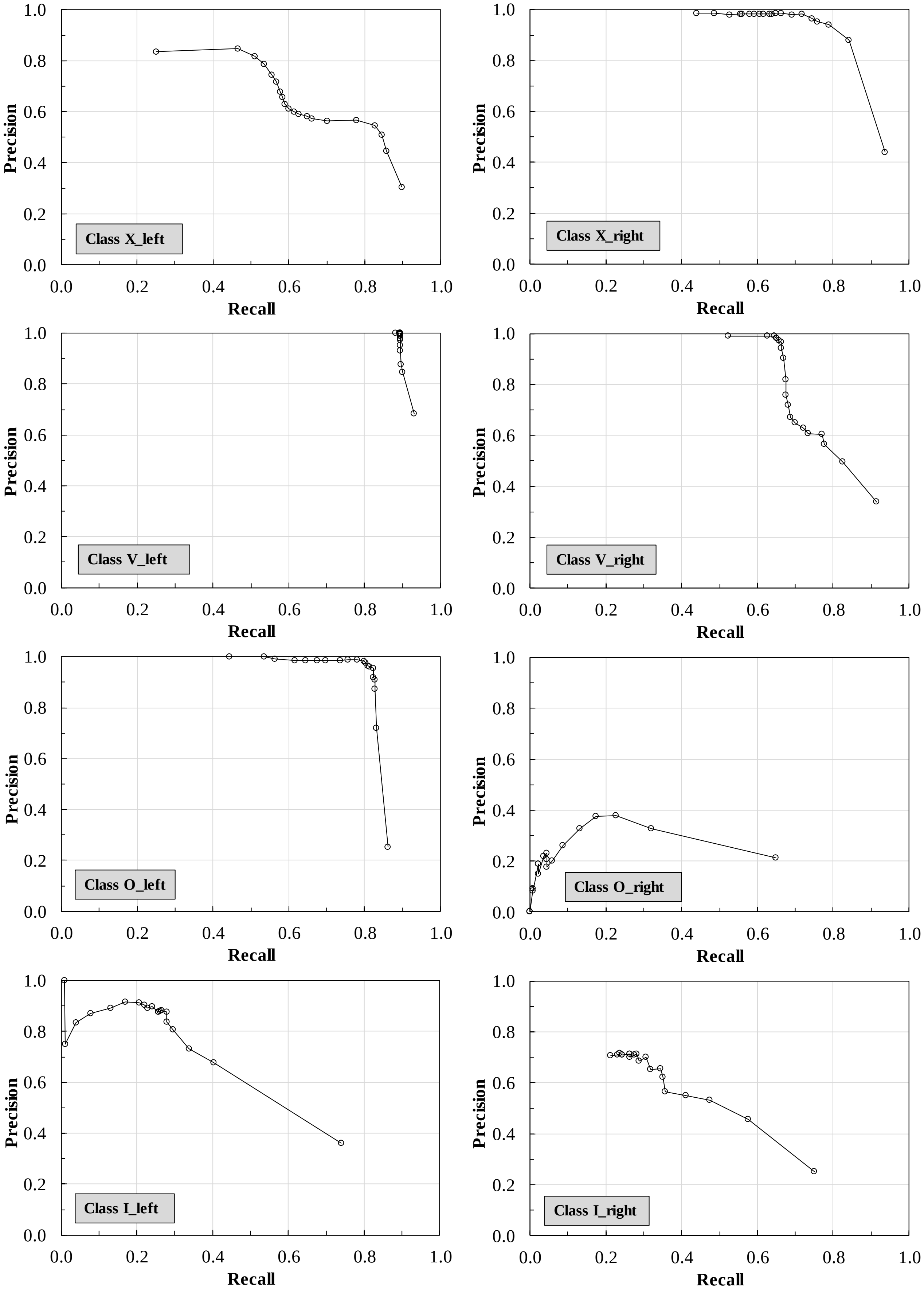


Figure 8. Precision-Recall diagram of the computer vision detection for each one of the 8 target classes by using the original data frames.

Even if the AP of the total dataset is 0.7356, some classes have AP value also rather small (i.e. O_{right}, I_{right} and I_{left} with AP values respectively of 0.17, 0.40 and 0.45). The low values for the three worst classes are confirmed by the unusual trends in Figure 8. This analysis should drive the future investigations, oriented them to define the main causes of the low scores and to look for the adequate corrective actions. Differently from the previous discussed scores, the AIoU values are quite similar for the various classes are the AIoU-based identification of the worst classes is more difficult. This confirms some well-known weaknesses and limitations of this metric (Szeliski, 2011) useful to decide, with regards to an object, whether a prediction is correct or not, but it is not suitable to describe the precision of the prediction. For the sake of completeness, the conducted validation test provides a value of mAP=0.6350 obtained as mean value among the 8 classes.

Finally, in order to establish possible correlations between the main features of the dataset and the outcomes from the validation test, the correlation matrix, reported in Figure 9, has been realized and adopted for the evaluation. Six independent variables, numerically quantifiable, have been selected (i.e. Occurrences, Average Area of the bounding boxes and the product Occurrences × Average Area of the bounding boxes for both training and validation datasets). Two dependent variables have been selected among the metrics adopted to evaluate the reliability of the detections (i.e. AP and AIoU).

Then the correlation matrix has dimension 8×8. The numerical values adopted for the creation of the correlation matrix are those reported in Table 2 and Table 4.

Table 4
Summary of the results from the validation test for each class by considering the original data frames.

	X_{left}	X_{right}	V_{left}	V_{right}	O_{left}	O_{right}	I_{left}	I_{right}	Total
AP	0.6485	0.8627	0.9202	0.7838	0.8336	0.1698	0.4583	0.4027	0.7356
AIoU	0.6625	0.7600	0.8547	0.7691	0.7553	0.5232	0.6999	0.7333	0.7812

The main aspects are the following:

- (see column #1, row #4,) the occurrences of the different classes populating training and validation datasets, shows good correlations confirming a proper subdivision of the frames into the training group and validation group (the slope of the linear regression represent the ratio 20%/80%=0.25 of used for the subdivision of the whole frame dataset);

- (see column #1, rows #7, #8 and #9) a characteristic trend exists between number of occurrences in the training dataset and the metrics used for the evaluation of the detection quality (i.e. AP and AIoU). By increasing the number of occurrences until a certain "threshold" value it increases in a considerable way the metric values, but after this threshold, the trend (see dashed red line in the subfigures) presents a knee characterized by a second branch with low slope. Then, it is like to say that after certain threshold (occurrences) value a considerable augment of the number of occurrences produces almost negligible improvement in the detections;
- (see column #4, rows #7, #8 and #9) the same evaluation is valid also for the relation between number of occurrences in the validation dataset and the metrics used for the evaluation of the detection quality;
- the area of the bounding boxes shows no significant correlation with the metrics of reliability of the detections (see columns #2 and #5). Nevertheless, the features obtained by the product "Occurrences × Average Area of the bounding boxes" (see columns #3 and #6) shows a positive correlation with the metrics, although it is not possible to identify a clear regression curve. However, if we exclude the two classes with values of the product $O_t \times A_t$ of the training phase significantly smaller than all the other classes, i.e. O_{left} and I_{right} , a quadratic regression curve of the relationship between AP and $O_t \times A_t$ is recognizable, with R^2=0.837 and the following equation:

$$y = -0.0272 x^2 + 0.7111 x - 3.8179 \quad (6)$$

This indicates that for the classes having $O_t \times A_t > 8$, AP rapidly increases up to $O_t \times A_t$ near to 12, then it becomes almost constant. This result shows that is possible to identify optimal values for the occurrence number and the bounding box areas in the training phase, which could be very useful to efficiently plan video acquisition to use for train the deep learning network. Therefore, this aspect deserves further and more in-depth investigations that will be carried out in future experimental campaigns carried out with additional video shooting.
- lastly (see columns #7, row #8) a positive correlation is confirmed between AP and AIoU.

These results provide useful indications for both the selection of the strategies more convenient in order to improve the detection quality and the development of optimal datasets for the application investigated in this paper.

First of all, if some classes are detected with poor precision, increasing the occurrences in the training dataset the detection accuracy is expected to rise. It makes no sense to increase the occurrences of all classes, especially for those classes that already have suitable metric values, because this would increase the costs of the labelling phase and the computational time of the training phase without producing considerable improvements.

Above a certain threshold value of the metrics, it also seems that increasing even the occurrences of the training dataset does not produce considerable benefits. So in such cases, probably alternative solutions are to be sought, such as the replacement of some data frames with others more informative for the network.

Finally, it seems that the average area of bounding boxes used for animal detection is not so influential, while the product of number of occurrences and box area is positively correlated with the average precision. This from a practical point of view has considerable advantages for the type of application investigated here, since typically in facilities such as cattle barns the videos are taken from considerable distance (even several tens of meters) due to logistical and security reasons of the cameras. On the other hand, the possibility to record videos from far away means that few cameras could be sufficient to cover large areas typical of cattle farms.

The evaluation and the development of these further steps in the process of developing the applications of computer vision systems to the dairy cow sector will be the first objectives of future research work.

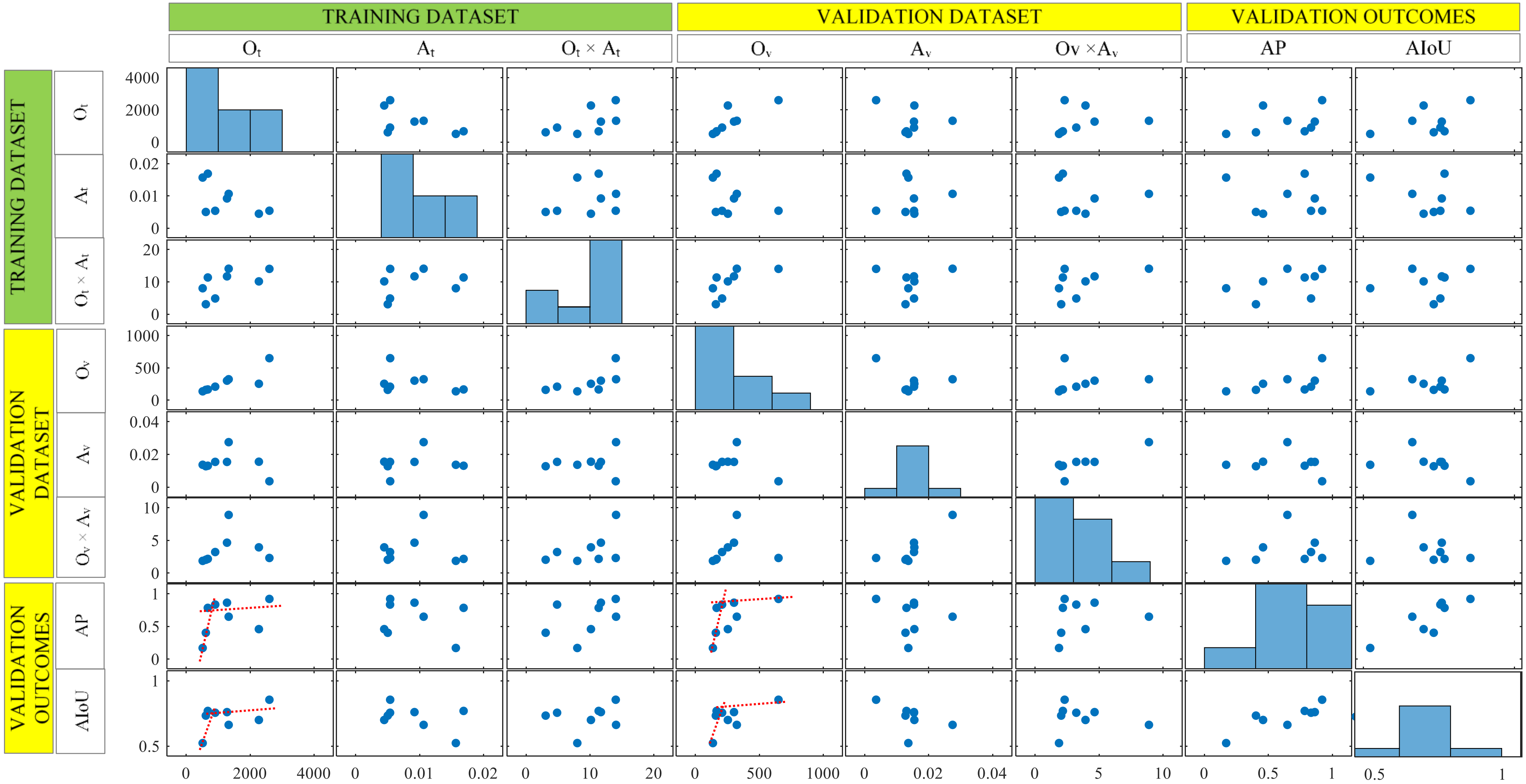


Figure 9. Correlation matrix between some features of the training dataset (i.e. O_t, A_t and $O_t \times A_t$) and validation dataset (i.e. O_v, A_v and $O_v \times A_v$) selected as independent variables and some metric outcomes of the validation phase selected as dependent variables.

3.2 *With augmented data frames*

This subsection presents a first preliminary attempt to increase the detection quality of some classes. In this case, the four classes with lowest total occurrence number (i.e. V_{right}, O_{left}, O_{right} and I_{right}) have been selected and by adopting a procedure of data augmentation, their total occurrence number have been increased. The two main objectives of the data augmentation test were:

i) to understand if simple alteration of the original frames can constitute a viable method to increase the available frame datasets in the context of cow detection and, following authors' knowledge, this represent the first application of this type of methods in the herd monitoring research field;

ii) to estimate the improvement of the detection performances of the network, in terms of AP, connected to the augment of the number of occurrences.

In order to judge the possible relation between augment of the number of occurrences and AP improvement, different threshold values have been investigated for the four classes. The following parameter Δ_O (%) has been adopted for the identification of the augment of the number of occurrences:

$$\Delta_O\,(\%) = \text{Augment}_t \,/\, O_t \times 100 \tag{7}$$

where: Augment_t is the is the increase of the occurrence number respect to the original dataset in the frames used for the training; O_t is the occurrence number in the frames of the original dataset used in the training. Four different target ranges have been considered for Δ_O (three below 50%, i.e. 10-20%; 20-30%; 30-40% to test the low rates of increase and one above, i.e. 60-80%) and every class has been associated to one of the ranges in order to increase the number of occurrences of the different classes in different ways. In this way has been possible to estimate a correlation between increase of occurrences and increase of detection performances when artificial frames are added in the dataset.

The data augmentation procedure has selected some frames, randomly extracted among those containing the four classes indicated above, and artificially have produced a modified copy of every selected frames. The modified copy has been obtained by changing the brightness level of the original frame so simulating possible different light conditions. This procedure, performed with the software XnConvert (Allan et al., 2019), creates a series of modified frames that could really occur in the stable. As an example, Figure 10 shows the comparison between the original and modified frame created by means of the described procedure. The 919 modified frames have been added to the original dataset (with 10105 frames) in order to constitute the augmented dataset (with 11024 frames in total).

(a) (b)

Figure 10. Comparison between (a) the original frame as recorded and (b) the modified frame created modifying the brightness of the image.

Also in this case the frames to be used in the validation phase were carefully selected in order to guarantee that all the 8 considered classes were adequately represented in the frames. The occurrences of each class, in the augmented dataset have been reported in Table 5, for both training and validation phases. The table also collects the augment of the occurrences, for each class, with respect to the original dataset. The augment of occurrences have generated Δ_O values equal to 29.0%, 15.9%, 71.3% and 37.5% for V_{right}, O_{left}, O_{right} and I_{right} respectively. In total, the augments of occurrences have been 1048, 158 and 1206 respectively for training, validation and total datasets. It is worth to highlight that original dataset, in terms of occurrences, has been augmented of about 10%, with the major increases related to training dataset of the classes V_{right}, O_{left}, O_{right} and I_{right}.

Table 5

Number of occurrences for each class for both training and validation augmented datasets. Augment is the increase on the occurrence number respect to the original datasets.

Training dataset									
	X_{left}	X_{right}	V_{left}	V_{right}	O_{left}	O_{right}	I_{left}	I_{right}	Sum
O_t	1383	1384	2668	868	1048	877	2147	840	11215
$Augment_t$	58	111	69	195	144	365	-123	229	1048
Validation dataset									
	X_{left}	X_{right}	V_{left}	V_{right}	O_{left}	O_{right}	I_{left}	I_{right}	Sum
O_v	331	302	656	176	218	139	377	161	2360
$Augment_v$	7	0	6	10	9	2	123	1	158

Then, the process follows the same steps used for the analysis on the original dataset, and for the sake of comparison with the previous case, analogous graphs and tables are reported in the following in order to summarize the main results. Figure 11(a) reports the global P-R graphs of the validation test obtained for both original and augmented datasets by considering all the 8 classes. From the comparison between the two curves it emerges that also the introduction of very similar frames, which differ only in brightness from the original, can improve both the precision (P) and the network's detection quality. In fact, for the present dataset, the precision improves in the CI range from 10% to 50% (see Figure 11(b)). Conversely, the recall (R) has an anti-symmetric trend with respect to CI of 50%. It improves in the CI from 0% to 50% and slightly deteriorates for CI higher than 50% (see Figure 11(c)). Same trend is obtained for the F_1-score (see Figure 11(d)). As far as the IoU metrics is concerned, it decreases about 8-9% all along the CI set. For the case at hand, the P value is adequate (i.e. higher than 0.8) for CI higher than 15%. For CI equal to 15% the R value is about 0.75. Table 6 collects all the results graphically reported in Figure 11.

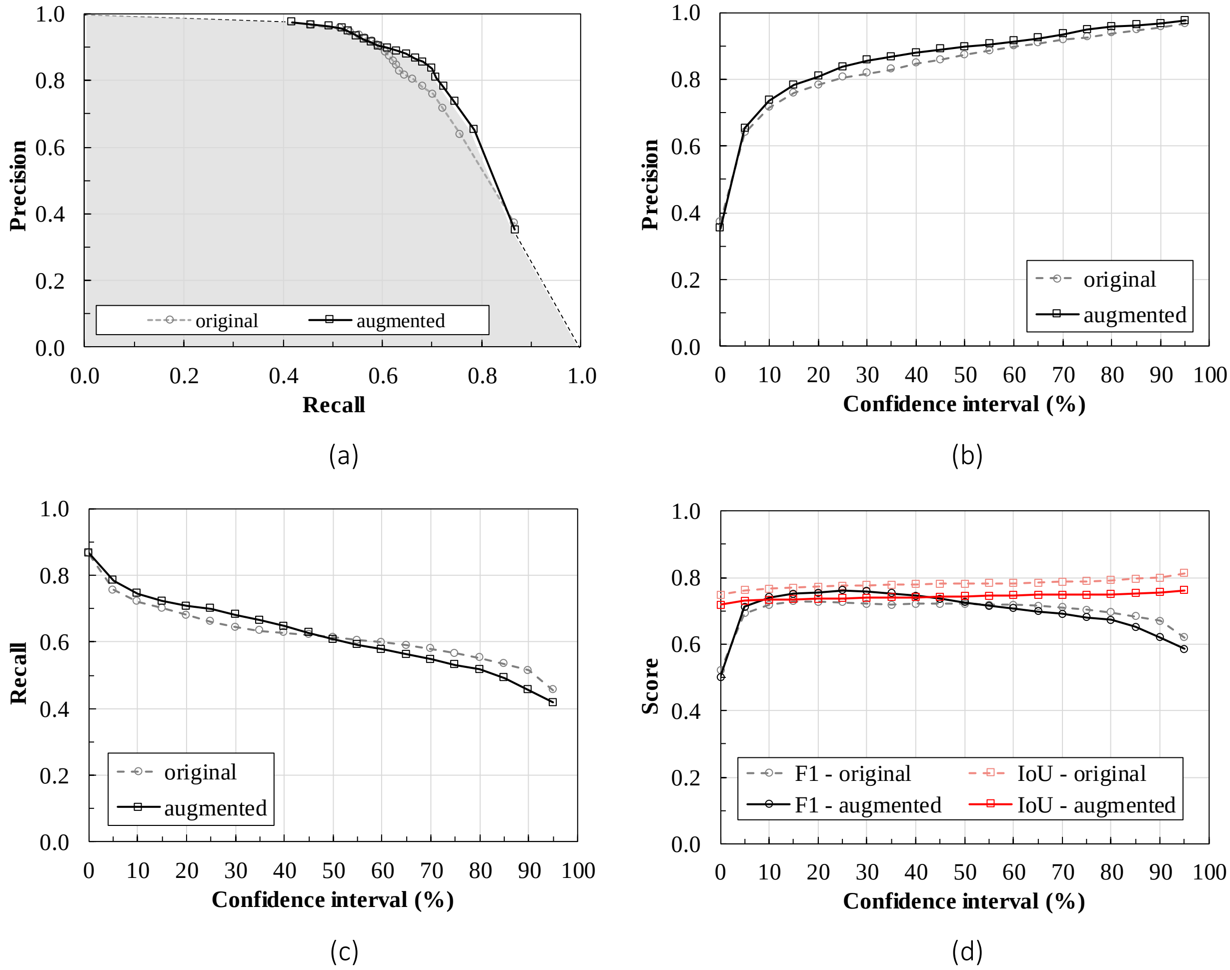


Figure 11. Comparison between the principal trends obtained from the validation test by considering all the occurrences dataset and reported for different confidence interval (with both original and augmented data frames). (a) P-R curve. (b) P trend; (c) R trend; (d) F_1 and IoU score for different confidence interval.

As far as the single class is concerned, in Table 7 are collected the parameters used to judge the detection quality of the various classes and Figure 12 shows the Precision-Recall graphs of each single class. From the table, the classes with better AP are V_{left}, O_{left}.and V_{right}. Instead, the classes with the worst AP are O_{right}, I_{right} and X_{left}.

Table 6

Main results from the validation test by considering all the occurrences dataset and reported for different confidence interval (with augmented data frames).

Confidence interval	True positive*	Ground truth**	Precision	Recall	F_1-score	IoU
0.95	986	1011	0.9753	0.4178	0.5850	0.7615
0.90	1076	1113	0.9668	0.4559	0.6196	0.7563
0.85	1162	1207	0.9627	0.4924	0.6515	0.7533
0.80	1223	1278	0.9570	0.5182	0.6723	0.7506
0.75	1253	1322	0.9478	0.5309	0.6806	0.7489
0.70	1293	1383	0.9349	0.5479	0.6909	0.7484
0.65	1328	1438	0.9235	0.5627	0.6993	0.7477
0.60	1363	1491	0.9142	0.5775	0.7079	0.7463
0.55	1396	1544	0.9041	0.5915	0.7152	0.7454
0.50	1437	1601	0.8976	0.6089	0.7256	0.7439
0.45	1483	1668	0.8891	0.6284	0.7363	0.7424
0.40	1529	1739	0.8792	0.6479	0.7460	0.7410
0.35	1571	1812	0.8670	0.6657	0.7531	0.7403
0.30	1609	1880	0.8559	0.6818	0.7590	0.7392
0.25	1652	1976	0.8360	0.7000	0.7620	0.7377
0.20	1671	2067	0.8084	0.7081	0.7549	0.7371
0.15	1707	2183	0.7820	0.7233	0.7515	0.7351
0.10	1760	2391	0.7361	0.7458	0.7409	0.7329
0.05	1854	2843	0.6521	0.7852	0.7125	0.7296
0.00	2090	5927	0.3526	0.8669	0.5013	0.7184

* : is the number of true positive occurrences detected from the neural network

** : is the number of “real” occurrences in the dataset as resulting from the visual detection performed by the operator.

Table 7

Summary of the results from the validation test for each class by considering the augmented data frames.

	X_{left}	X_{right}	V_{left}	V_{right}	O_{left}	O_{right}	I_{left}	I_{right}	Total
AP	0.5489	0.8755	0.9405	0.8815	0.8954	0.2405	0.7618	0.5391	0.7559
AIoU	0.7034	0.7593	0.7481	0.8153	0.7221	0.6392	0.6635	0.7865	0.7428

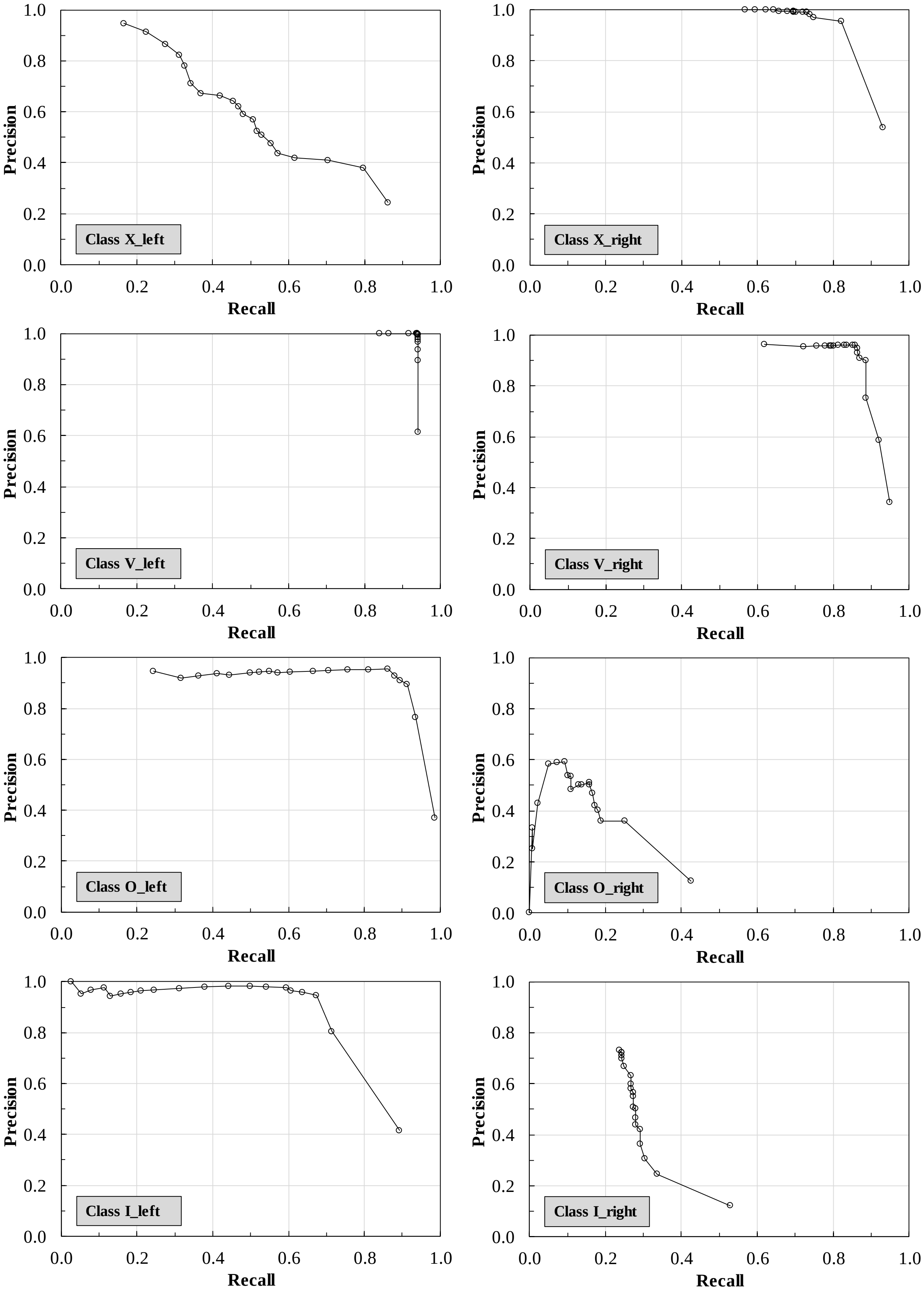


Figure 12. Precision-Recall diagram of the computer vision detection for each one of the 8 target classes by using the augmented data frames.

In order to compare the network detection performances of the original and augmented data frame cases, the following parameter Δ_{AP} (%) has been defined and calculated for the four classes majorly influenced by the data augmentation procedure:

$$\Delta_{AP}\ (\%) = (AP_{augmented} - AP_{original}) / AP_{original} \times 100 \tag{8}$$

where: $AP_{augmented}$ is the average precision obtained using the augmented dataset; $AP_{original}$ is the average precision obtained using the original dataset.

The values of Δ_{O} and Δ_{AP}, representing a (percentage) relative difference of occurrences and average precision respectively, are reported in Table 8 for the four classes of interest.

Table 8

Percentage relative difference of occurrences (Δ_{O}) and average precision (Δ_{AP}) for the four classes mainly influenced by the data augmentation procedure.

	V_{right}	O_{left}	O_{right}	I_{right}
Δ_{O} (%)	29.0	15.9	71.3	37.5
Δ_{AP} (%)	12.5	7.4	41.6	33.9

By the analysis of the Δ_{AP} values in Table 8, it emerges that the augmented dataset is able, in general, to provide an improvement of the metric values, or i.e. is able to increase also the single class detection quality of the neural network, especially for those classes characterized by a considerable augment of occurrences. This was probably expected, but the most important outcomes may be that, also for the single classes, the introduction of artificially-obtained frames is suitable to increase the detection scores. This aspect has important practical implications since in applications similar to the one studied here it is not always possible to record videos for long periods.

Figure 13 graphically shows the position of the values of Δ_{O} and Δ_{AP} reported in Table 8. The further emerging outcome is that exist a clear direct relation between the increase of the occurrences number and the detection performances of the network and the relation seems almost linear, at least for the investigated ranges.

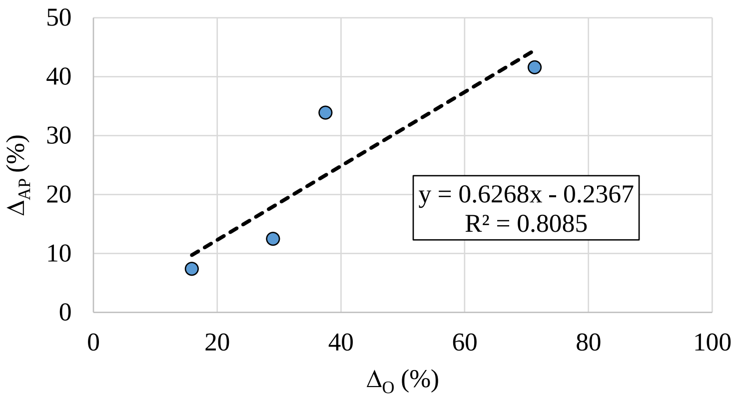


Figure 13. Percentage relative difference of occurrences (Δ_O) Vs. average precision (Δ_{AP}) for the four classes mainly influenced by the data augmentation procedure.

These outcomes confirm that for the situations in which the number of frames is not sufficient to provide a suitable occurrence number for some classes, in order to increase the detection performance of the network, a reliable strategy could be augment the dataset by adopting “virtual” frames made in-house and ad-hoc for the single classes not adequately represented in the acquired videos. This will be object of future investigations and further details are not reported here since they are beyond the scope of this paper. For the sake of completeness, the conducted validation test provides a value of mAP=0.6604, higher than the value obtained for the original dataset.

The developed system is suitable to implement various future extensions performing tracking and identification of anomalous behaviours. Cows tracking can be achieved by the integration of algorithms capable to compute the IoU between chronologically subsequent frames. IoU rate can be interpreted as displacement of the animal if it overcomes a predefined threshold. This will allow to record the trajectories of individual animals and also to compute the time spent in different positions, thus providing an accurate proxy of the time budget of every cow. Furthermore, this approach can be also used to detect standing time and lying bouts, which can be used to assess welfare indices of the herd, or different groups or even individual animals. Moreover, the evaluation of a different detection algorithm will be also object of future investigations.

4. Conclusions

This study represents the first step for the development of a detection system aiming to recognize individual cows, evaluate their position, understand the action the cow is carrying out and finally tracking the cow movements in the barn. A computer vision system based on deep learning models for the automatic detection of individual cow based on the pelt pattern, within images using an HD resolution camera, was designed and implemented in a case study barn.

The global detection performances of the network, reported in terms of precision-recall curves, have been proved to be good for some classes and excellent for others since the AIoU is about 0.78 and the IoU for the different classes ranges from 0.75 to 0.81. The performances of the network is confirmed by the F_1-score ranging from 0.67 to 0.73 for common confidence interval from 5% to 90%. The outcomes proved that the natural pattern of the cow pelt investigated in the study is suitable for the animal detection, all this representing a necessary step prior to understand the cow action. Moreover, in this study a useful still simple equation has been proposed for the evaluation of the optimal values for the occurrence number and the bounding box areas in the training phase. The regression procedure for the equation calibration showed that a quadratic relation exists between AP and O_t x A_t. This proposal could be very useful to efficiently plan the acquisition to use for train the network in this type of context. Finally, the application of a very simple data augmentation technique, changing the frame brightness level, has been confirmed to be an effective strategy to adopt in order to improve the performances of the network, in case of insufficient occurrences.

The promising results reported here present the first phase of the work for the definition of a computer vision-based system for herd monitoring applications devoted to the study of movements, actions and behavior of the cows in a barn.

Acknowledgements

The authors wish to thank Prof. Andrea Formigoni, Scientific Supervisor of the Experimental and Didactic Dairy Cows Unit of the University of Bologna, where all the video recordings, images and data necessary for carrying out the research have been acquired.

Funding

The activity presented in the paper is part of the research project PRIN 2017 "Smart dairy farming: innovative solutions to improve herd productivity" funded by the Italian Ministry of Education, University and Research [20178AN8NC].

References

AFIMILK, 2020. Cow monitoring. website. URL https://www.afimilk.com/cow-monitoring (accessed 5.15.20).

Allan, E.L., Livermore, L., Price, B.W., Shchedrina, O., Smith, V.S., 2019. A Novel Automated Mass Digitisation Workflow for Natural History Microscope Slides. Biodivers. data J. 7, e32342–e32342. https://doi.org/10.3897/BDJ.7.e32342

ALLFLEX, 2020. Herd monitoring. website. URL https://westgen.com/products/scr-herd-monitoring (accessed 5.20.20).

Alsaaod, M., Fadul, M., Steiner, A., 2019. Automatic lameness detection in cattle. Vet. J. https://doi.org/10.1016/j.tvjl.2019.01.005

Aydin, A., 2017. Development of an early detection system for lameness of broilers using computer vision. Comput. Electron. Agric. 136, 140–146. https://doi.org/10.1016/j.compag.2017.02.019

Barkema, H.W., von Keyserlingk, M.A.G., Kastelic, J.P., Lam, T.J.G.M., Luby, C., Roy, J.-P., LeBlanc, S.J., Keefe, G.P., Kelton, D.F., 2015. Invited review: Changes in the dairy industry affecting dairy cattle health and welfare. J. Dairy Sci. 98, 7426–7445.

Berckmans, D., 2014. Precision livestock farming technologies for welfare management in intensive livestock systems. Rev. sci. tech. Off. int. Epiz 33, 189–196. https://doi.org/10.20506/rst.33.1.2273

Bewley, J.M., Robertson, L.M., Eckelkamp, E.A., 2017. A 100-Year Review: Lactating dairy cattle housing management. J. Dairy Sci. https://doi.org/10.3168/jds.2017-13251

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

Cowley, F.C., Barber, D.G., Houlihan, A. V, Poppi, D.P., 2015. Immediate and residual effects of heat stress and restricted intake on milk protein and casein composition and energy metabolism. J. Dairy Sci. 98, 2356–68. https://doi.org/10.3168/jds.2014-8442

DELAVAL, 2020. Sensors for herd. website. URL https://www.delaval.com (accessed 5.12.20).

Fournel, S., Rousseau, A.N., Laberge, B., 2017. Rethinking environment control strategy of confined animal housing systems through precision livestock farming. Biosyst. Eng. 155, 96–123. https://doi.org/10.1016/j.biosystemseng.2016.12.005

Girshick, R., 2015. Fast R-CNN, in: Proceedings of the IEEE International Conference on Computer Vision. pp. 1440–1448. https://doi.org/10.1109/ICCV.2015.169

Girshick, R., Donahue, J., Darrell, T., Malik, J., 2014. Rich feature hierarchies for accurate object detection and semantic segmentation, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. pp. 580–587. https://doi.org/10.1109/CVPR.2014.81

Guzhva, O., Ardö, H., Herlin, A., Nilsson, M., Åström, K., Bergsten, C., 2016. Feasibility study for the implementation of an automatic system for the detection of social interactions in the waiting area of automatic milking stations by using a video surveillance system. Comput. Electron. Agric. 127, 506–509. https://doi.org/10.1016/j.compag.2016.07.010

Halachmi, I., Klopčič, M., Polak, P., Roberts, D.J., Bewley, J.M., 2013. Automatic assessment of dairy cattle body condition score using thermal imaging. Comput. Electron. Agric. 99, 35–40. https://doi.org/10.1016/j.compag.2013.08.012

Jaeger, M., Brügemann, K., Brandt, H., König, S., 2019. Associations between precision sensor data with productivity, health and welfare indicator traits in native black and white dual-purpose cattle under grazing conditions. Appl. Anim. Behav. Sci. https://doi.org/10.1016/j.applanim.2019.01.008

Jiang, B., Song, H., He, D., 2019. Lameness detection of dairy cows based on a double normal background statistical model. Comput. Electron. Agric. 158, 140–149. https://doi.org/10.1016/j.compag.2019.01.025

Kamilaris, A., Prenafeta-Boldú, F.X., 2018. Deep learning in agriculture: A survey. Comput. Electron. Agric. https://doi.org/10.1016/j.compag.2018.02.016

Li, W., Ji, Z., Wang, L., Sun, C., Yang, X., 2017. Automatic individual identification of Holstein dairy cows using tailhead images. Comput. Electron. Agric. 142, 622–631. https://doi.org/10.1016/j.compag.2017.10.029

Martinez-Ortiz, C.A., Everson, R.M., Mottram, T., 2013. Video tracking of dairy cows for assessing mobility scores, in: Joint European Conference on Precision Livestock Farming,. 10 – 12 September 2013, Leuven, Belgium, p. 8.

Microsoft, 2018. VoTT: Visual Object Tagging Tool. GitHub Repos.

Nie, X., Yang, M., Liu, R.W., 2019. Deep Neural Network-Based Robust Ship Detection Under Different Weather Conditions, in: 2019 IEEE Intelligent Transportation Systems Conference (ITSC). pp. 47–52. https://doi.org/10.1109/ITSC.2019.8917475

Norouzzadeh, M.S., Nguyen, A., Kosmala, M., Swanson, A., Palmer, M., Packer, C., Clune, J., 2017. Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning, in: Proceedings of the National Academy of Sciences of the United States of America. pp. 1–17. https://doi.org/10.1073/pnas.1719367115

Okura, F., Ikuma, S., Makihara, Y., Muramatsu, D., Nakada, K., Yagi, Y., 2019. RGB-D video-based individual identification of dairy cows using gait and texture analyses. Comput. Electron. Agric. 165, 104944. https://doi.org/10.1016/j.compag.2019.104944

Porto, S.M.C., 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. Biosyst. Eng. 133, 46–55. https://doi.org/10.1016/j.biosystemseng.2015.02.012

Porto, S.M.C., Arcidiacono, C., Anguzza, U., Cascone, G., 2013. A computer vision-based system for the automatic detection of lying behaviour of dairy cows in free-stall barns. Biosyst. Eng. 115, 184–194. https://doi.org/10.1016/j.biosystemseng.2013.03.002

Redmon, J., 2013. 2016. Darknet: Open Source Neural Networks in C [WWW Document]. website. URL https://pjreddie.com/darknet/ (accessed 5.10.20).

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. IEEE, pp. 779–788. https://doi.org/10.1109/CVPR.2016.91

Redmon, J., Farhadi, A., 2018. YOLOv3: An Incremental Improvement. arXiv Prepr. 1804.02767.

Ren, S., He, K., Girshick, R., Sun, J., 2017. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. IEEE Trans. Pattern Anal. Mach. Intell. 39, 1137–1149. https://doi.org/10.1109/TPAMI.2016.2577031

Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., Savarese, S., 2019. Generalized intersection over union: A metric and a loss for bounding box regression. Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. 2019-June, 658–666. https://doi.org/10.1109/CVPR.2019.00075

Song, X., Leroy, T., Vranken, E., Maertens, W., Sonck, B., Berckmans, D., 2008. Automatic detection of lameness in dairy cattle-Vision-based trackway analysis in cow's locomotion. Comput. Electron. Agric. 64, 39–44. https://doi.org/10.1016/j.compag.2008.05.016

Szeliski, R., 2011. Computer Vision, I. ed. Springer-Verlag London, London. https://doi.org/10.1007/978-1-84882-935-0

Taigman, Y., Ming, Y., Ranzato, M., Wolf, L., 2014. DeepFace: Closing the Gap to Human-Level Performance in Face Verification, in: Conference on Computer Vision and Pattern Recognition (CVPR). https://doi.org/10.1109/CVPR.2014.220

Trnovszky, T., Kamencay, P., Orjesek, R., Benco, M., Sykora, P., 2017. Animal recognition system based on convolutional neural network. Adv. Electr. Electron. Eng. 15, 517–525. https://doi.org/10.15598/aeee.v15i3.2202

Tsai, D.M., Huang, C.Y., 2014. A motion and image analysis method for automatic detection of estrus and mating behavior in cattle. Comput. Electron. Agric. 104, 25–31. https://doi.org/10.1016/j.compag.2014.03.003

Tullo, E., Finzi, A., Guarino, M., 2019. Review: Environmental impact of livestock farming and Precision Livestock Farming as a mitigation strategy. Sci. Total Environ. https://doi.org/10.1016/j.scitotenv.2018.10.018

Van Hertem, T., Alchanatis, V., Antler, A., Maltz, E., Halachmi, I., Schlageter-Tello, A., Lokhorst, C., Viazzi, S., Romanini, C.E.B., Pluk, A., Bahr, C., Berckmans, D., 2013. Comparison of segmentation algorithms for cow contour extraction from natural barn background in side view images. Comput. Electron. Agric. 91, 65–74. https://doi.org/10.1016/j.compag.2012.12.003

Van Hertem, T., Schlageter Tello, A., Viazzi, S., Steensels, M., Bahr, C., Romanini, C.E.B., Lokhorst, K., Maltz, E., Halachmi, I., Berckmans, D., 2018. Implementation of an automatic 3D vision monitor for dairy cow locomotion in a commercial farm. Biosyst. Eng. https://doi.org/10.1016/j.biosystemseng.2017.08.011

Van Hertem, T., Steensels, M., Viazzi, S., Romanini, E.C.B., Bahr, C., Berckmans, D., Schlageter Tello, A., Lokhorst, K., Maltz, E., Halachmi, I., 2014. Improving a computer vision lameness detection system by adding behaviour and performance measures, in: International Conference of Agricultural Engineering. pp. 1–8.

Ventura, B.A., Von Keyserlingk, M.A.G., Wittman, H., Weary, D.M., 2016. What difference does a visit make? Changes in animal welfare perceptions after interested citizens tour a dairy farm. PLoS One. https://doi.org/10.1371/journal.pone.0154733

CRediT author statement

Patrizia Tassinari: Conceptualization, Funding acquisition, Project administration, Resources, Supervision.

Marco Bovo: Conceptualization, Data curation, Formal analysis, Methodology, Validation, Visualization, Writing—original draft, Writing—review & editing.

Stefano Benni: Conceptualization, Formal analysis, Investigation, Methodology, Supervision, Validation, Visualization, Writing—original draft.

Simone Franzoni: Software

Matteo Poggi: Software

Ludovica Maria Eugenia Mammi: Investigation

Stefano Mattoccia: Conceptualization, Software, Validation

Luigi Di Stefano: Conceptualization, Software, Validation

Filippo Bonora: Data curation

Alberto Barbaresi: Visualization, Writing—review & editing

Enrica Santolini: Visualization, Writing—review & editing

Daniele Torreggiani: Conceptualization, Project administration, Resources, Supervision, Writing—review & editing.