A CNN-based Cow Interaction Watchdog

Håkan Ardö, Oleksiy Guzhva, Mikael Nilsson

September 30, 2016

Abstract

Animal behaviour and welfare can be studied/assessed by looking at different interactions occurring between the animals. Video recordings of a scene of interest are often made and then watched/evaluated by experts. However, the interactions of interest are often fairly rare. To reduce the amount of time the experts spend on watching the uninteresting video, this paper introduces an automated watchdog system that can discard some of the recorded video material. A pilot study on cows was made where a Convolutional Neural Network (CNN) detector was used to count the number of cows in the scene and discard video where less than two cows were present. This removed 38 % of the recordings while only losing 1 % of the interesting video.

1 Introduction

Scientists working with animal behaviour and welfare are interested in studying the social interactions between cows in dairy farms. Typically these studies performed by defining a set of interactions such as head butting, body pushing, social licking etc. and writing a very detailed protocol with the description of every interaction. Then an expert studies the area of interest for a large amount of time and counts the number of each interaction occurring for the whole duration of the video sequence/sequences used for the particular study [6]. Some of these interactions are quite rare, which means that a lot of expert time have to be spent in looking at raw video data in order to find potentially interesting sequences.

A number of recent studies on cow behaviour in the dairy barn environment were based on different GPS or wireless sensor network (WSN) solutions [7]. These products allow scientists to see spatial distribution of animals and to measure different levels of activity [7]. However, the number of behaviours that could be monitored with the position-based approach is limited. Therefore, new methods based on video surveillance and image analysis, which could extend the number of parameters for studying, are of great importance.

In this paper, the goal is to take the first step towards an automated system for behavioural analysis. The study area is filmed using video cameras. Then an automated watchdog system will remove irrelevant parts (e.g. those, containing events that are not relevant or without animals in the scene) of the recorded video material. The remaining video sequences will still have to be studied by experts, but the time spent looking at uninteresting video sequences will be significantly reduced.

The pilot study used to develop this watchdog was made in a dairy barn in the south of Sweden with 252 Swedish Holstein cows. Cows were milked by four automatic milking robots, which had a common waiting area (6x18 meters). This waiting area is a common space which cows that are ready for milking could enter at any point in time. They will then interact with each other in order to decide who are allowed to enter each of the milking robots and in which order. A direct relationship between cows' inter-cow distance and their aggressive behavioural patterns was demonstrated by [1]. Other studies [3, 5] showed inevitable effects of inferior animal welfare connected to the restrained performance of their natural behaviour. Therefore, early diagnostics of unconditional changes in animal behaviour when linked to health and welfare could not only save time and money for the farmer but also decrease the production pressure for every animal in the barn [8].

1.1 Experimental setup

Video recordings were made using three Axis M3006-V cameras with a wide angle of 134 degrees that were placed at the 3.6-meter height, pointing straight down to optimise overview over the study area. There is a significant overlap between the camera images in order to not miss events taking place at the border between the cameras. In total 2315 hours (1 month) of 800x600 video in 16 Frames Per Second (FPS), was collected.

The cameras were calibrated to compensate for lens distortion and rectified. Although the cameras were physically mounted to point fairly straight down, they were still slightly tilted. This tilting was synthetically removed during the rectification. The end result of this calibration is video images where the cows have the same size regardless of where in the image they appear. Also, the scanlines of the three different cameras become aligned, which allows them to be stitched together to form an overview of the entire waiting area.

Finally, a Convolutional Neural Network (CNN) was trained to detect the cows in the images, and statistics about how many cows and their distances/relation to each other was extracted. Using that statistical data, scientists working in a field of animal behaviour could form queries to select particular time intervals to watch, such as "show me video clips involving at least two cows with the neck of one cow closer than one meter to the body of the other".

2 Camera calibration

Straight lines were manually annotated in the camera images. Focal length and lens distortion parameters for the camera model used in OpenCV version 2.4.9.1 were then optimised until the projections of the lines were straight in the images. Some of these lines, typically from the walls, were long enough to pass through all three cameras. These lines were used to find the orientation of the cameras, again by local optimisation until the projections of the lines into the different camera images agree.

3 CNN cow detector

A random subset of the full recording consisting of 1722 images was manually annotated. This subset contained in total 6399 cows. Each cow was annotated with seven landmark points: head, left and right shoulder, front middle, left and right hip and back middle. In addition to that one additional landmark "cow centre" was defined as the mean of front middle and back middle. This data was then used to train a CNN detector.

The detector was split into two steps. The first step is a fully convolutional CNN that detects the landmarks in the image. Currently, only four of the landmarks were used to speed up the experiments, but extending to use all seven is straightforward. The architecture of this network is a fully convolutional version of VGG [9] with batch normalisation [4] after each convolution step. Details are shown in Table 1.

The second step is another CNN that works with the probability map produced by the first as input and tries to detect the cows and their orientations. The full circle is divided into 32 equally spaced orientations which generate 32 different oriented cow classes. In addition to that, there is the no cow class, which makes the total number of classes of this CNN 33. The input probabilities were turned into log likelihoods as it makes more sense when summing them together. Then the network consists of a single 13×13 convolutional layer. Details are shown in Table 2.

The landmark net was trained on patches of 150×150 pixels extracted from the input images. This makes the output during training a single pixel. The positive examples were centred on the landmarks and randomly jittered ± 16 pixels (as the distance between output pixels is 32 input pixels). Negative patches were selected at centres more than 32 pixels from any landmark. In addition to the positive and negative patches a set of do not care patches were selected at random centres at distances between 16 and 32 pixels from landmarks. The ground truth probability of these patches belong to the class of the landmark was set to 0.5 and the probability that they are ground was set to 0.5. In some cases, several landmarks appear within 32 pixels of the patch centre. In that case, the probability mass was distributed uniformly among all involved classes.

The weights of the convolutions are initiated using random samples draw from a Gaussian distribution truncated at 2σ, with standard deviation

Layer type	Size	Channels
Conv + BNorm + Relu	3x3	32
MaxPool(stride=2)	2x2	
Conv + BNorm + Relu	3x3	64
MaxPool(stride=2)	2x2	
Conv + BNorm + Relu	3x3	128
Conv + BNorm + Relu	3x3	128
MaxPool(stride=2)	2x2	
Conv + BNorm + Relu	3x3	256
Conv + BNorm + Relu	3x3	256
MaxPool(stride=2)	2x2	
Conv + BNorm + Relu	3x3	512
Conv + BNorm + Relu	3x3	512
MaxPool(stride=2)	2x2	
Conv + BNorm + Relu	1x1	1024
Conv + BNorm + Relu	1x1	1024
Conv + BNorm + Relu	1x1	5
Softmax		

Table 1: CNN architecture for landmark detection.

Layer type	Size	Channels
MaxPool(stride=1)	3x3	
Log		
Conv + BNorm + Relu	13x13	33
Softmax		

Table 2: CNN architecture to detect oriented cows

$\sigma = \sqrt{\frac{2}{n}}$, where n is the number of inputs[2]. The networks are regularised with weight decay of 0.0001 and optimised using stochastic gradient descent with 0.9 momentum. The learning rate is initiated to 1.0 and reduced by a factor 10 each time the validation error flattens. The landmark CNN uses only valid outputs from the convolutional and maxpool layers while the cow detector keeps the same resolution to also detect cows that are slightly outside the image.

Once the net was trained, the last maxpool layer was removed to increase the output resolution. The net was then applied to the full rectified training images producing probability maps of $44 \times 46 \times 5$ pixels. These were used as training examples for the cow detection net (without splitting them into patches). Output ground truth probability maps of $44 \times 46 \times 33$ pixels were constructed from the annotations by projecting each cow, i, center point into the probability map as (x_i, y_i) and calculate its angle a_i as the angle of the line between front middle and back middle landmarks. Then a binary $44 \times 46 \times 33$ mask $B(x, y, c)$ is formed, containing a background mask

$$B(x, y, 32) = \begin{cases} 0 & \text{if} \quad \lfloor x_i \rfloor \leq x \leq \lceil x_i \rceil \\ & \quad \lfloor y_i \rfloor \leq y \leq \lceil y_i \rceil \\ 1 & \text{otherwise} \end{cases} \tag{1}$$

and 32 orientation masks

$$B(x, y, c) = \begin{cases} & \lfloor x_i \rfloor - 1 \leq x \leq \lceil x_i \rceil + 1 \\ 1 & \text{if} \quad \lfloor y_i \rfloor - 1 \leq y \leq \lceil y_i \rceil + 1 \\ & \quad \text{adist}\left(\frac{2c\pi}{32}, c_i\right) < \frac{2\pi}{32} \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

for $0 \leq c \leq 31$ and all i. The adist function calculates the absolute angular distance between two angles. The ground truth probability masks are then produced by normalising B to sum to 1 for each pixel. Finally, the network is trained using the same hyper parameters as described above.

4 Watchdog evaluation

To evaluate the system, the 6400 frames spread over the entire recording were processed by the CNN. A simple watchdog extracting frames containing two or more cows were implemented. That would be the most basic requirement for an interaction, and already with this simple criteria, it was possible to discard 38 % of the recordings as uninteresting. 50 random frames selected by the watchdog and 50 random frames discarded by the watchdog were automatically annotated by using the CNN results and studied manually. Cows intersecting the borders were ignored in the sense that the images were considered correct regardless of whether such border cases was detected or not.

97 % of the images were perfectly interpreted, i.e. all cows present were detected and no extra detections. Two of the images with errors containing several detected cows and was thus correctly classified as containing two or more cows by the watchdog, resulting in a watchdog hit rate for it of 99 %. In total, those 100 images contain 222 cows. One of those were not detected and 2 extra detections were made yielding a cow hit rate of 99.6 % with a

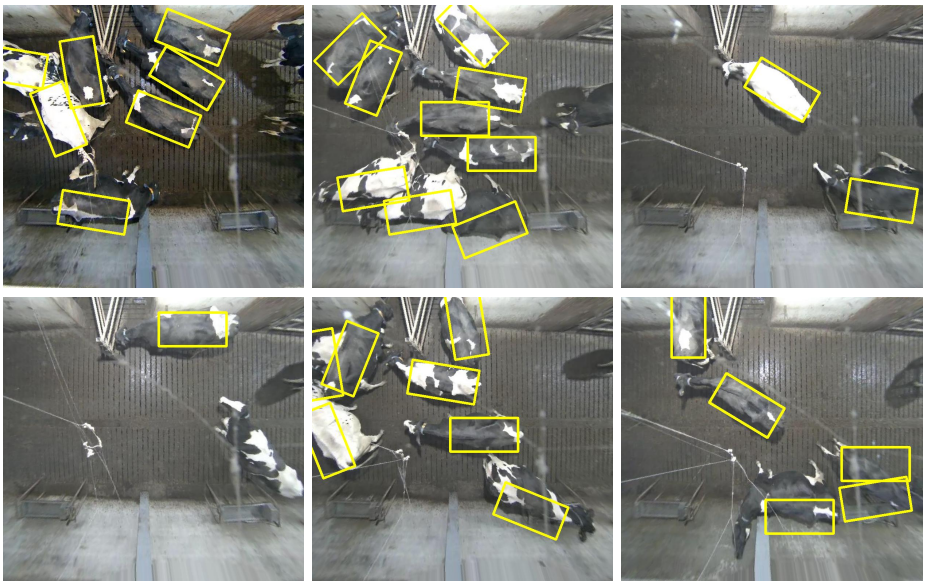

Figure 1: Top row: Three images correctly interpreted (all cows detected and no extra detections). Bottom row: The three images were the errors were made (1 missed cow and two extra detections).

false alarm rate of 0.9%. Some example detections are shown in Figure 1. Two of the reasons for mistakes are inter-cow occlusion and the combination of landmarks from different individuals.

The evaluation runs at 6.55 fps on a single Tesla K20m GPU, using single precision floats.

5 Conclusions

A CNN cow detection system has been developed. It can detect and count the cows present in the image with high precision. 97 % of the test images were perfectly interpreted in the sense that the system was able to place a rotated rectangle on each cow and nowhere else, c.f. Figure 1. This detector was used to discard 38 % of the recorded video as uninteresting while only losing 1 % of the interesting video.

References

[1] T. J. DeVries, M. A. G. von Keyserlingk, and D. M. Weary. Effect of feeding space on the inter-cow distance. *Aggression, and Feeding Behavior of Free-Stall Housed Lactating Dairy Cows. J. Dairy Sci*, 87, 2004.

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. *CoRR*, abs/1502.01852, 2015.

[3] P.. Hemsworth. Human-animal interactions in livestock production. *Appl. Anim. Behav. Sci*, 81, 2003.

[4] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *CoRR*, abs/1502.03167, 2015.

[5] R. J. Kilgour. In pursuit of 'normal': A review of the behaviour of cattle at pasture. *Appl. Anim. Behav. Sci*, 138, 2012.

[6] P. Martin and P. Bateson. *Measuring Behaviour An Introductory Guide*. Cambridge University Press, Cambridge, 2007.

[7] E. S. Nadimi, R. N. Jörgensen, V. Blanes-Vidal, and S. Christensen. Monitoring and classifying animal behavior using ZigBee-based mobile ad hoc wireless sensor networks and artificial neural networks. *Comput. Electron. Agric*, 82, 2012.

[8] A. Polikarpus, T. Kaart, H. Mootse, De Rosa, Arney G., and D. Influences of various factors on cows' entrance order into the milking parlour. *Appl. Anim. Behav. Sci*, 166, 2015.

[9] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.