# A CNN-based Cow Interaction Watchdog

Håkan Ardö[1*], Oleksiy Guzhva[2], Mikael Nilsson[1]

[1]Centre for Mathematical Sciences, Lund University, Sölvegatan 18, Lund, Sweden
[2]SLU, XXX
*ardo@maths.lth.se

**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. If also the distance between the cows present in the scene was considered, then 50 % of the recordings could be remoed without loosing more than 4 % of the interesting frames.

## 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 significant 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 [10]. Some of these interactions are relatively rare, which means that much expert time has to be spent in looking at raw video data to find potentially interesting sequences.

Some recent studies on cow behaviour in the dairy barn environment were based on different GPS or wireless sensor network (WSN) solutions [12]. These solutions allow scientists to see spatial distribution of animals and to measure varying levels of activity [12]. However, the number and complexity 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 unnecessary 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. This is achieved using a cow detector that detects cows in the scene and represents them with rotated bounding boxes. That means the number of cows can be counted and distances

between the cows can be measured. Rotated bounding boxes allow much more precise distance measures between diagonally oriented cows as compared to more classical non-rotated bounding boxes due to the elongated aspect ratio of the cows.

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 are allowed to 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 [2]. Other studies [6, 8] showed certain 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 [15].

## 1.1. Related work

Most previous work on detecting cows by using video cameras have been focused on monitoring areas where the orientation of the cows was known due to physical limitations imposed by the surroundings. Two examples of this are the Viola-Jones based detector of Arcidiacono et al. [20] for detecting cows at the feed barrier and the work of Martinez-Ortiz1 et al. [11] to detect and track cow heads in narrow entrance corridors. There are also approaches that rely on more advanced sensors, such as Abdul et. al [1] that uses a depth camera.

The current state of the art for detecting cows freely moving around was presented by Porto et al. [16]. They use a Viola-Jones based detector and need 6 cameras at 4.6 meters height to cover a $15.4 \times 3.8$ m area and detects cows in 3 different orientations: vertical, horizontal and diagonal with a hit rate of 90%. They do not use separate detectors for the two different diagonals which mean that producing rotated bounding box from their results would not be straight forward. In the work presented in this paper, only 3 cameras at 3.6 meters height were needed to cover a $18 \times 6$ m area and detect cows in 32 different orientations with a hit rate of 97%. Note that the hit rates are from different datasets, and that the dataset used in this article have larger variations in viewpoints due to the use of fewer cameras at a lower height to cover a larger area.

General purpose object detection frameworks such as YOLO [17, 18] and SSD [9] have very good performance. They do however focus on detecting objects of varying size and aspect ratio but with fix orientation. In the scenario considered in this paper, the size and aspect ratio is fixed and known while orientation (rotation) varies and have to be estimated.

He et al. [5] have considered rotated bounding boxes to generate general purpose objected proposals. They do a local optimization over position, size, aspect ratio and rotation from an initial set of sampled widows. In the proposed approach the size and aspect ratio are known and fixed and an exhaustive search over position and rotation is performed. That way there the risk of not considering relevant rotations is eliminated. Also, He et al. [5] is based on hand crafted features assuming that image borders are unlikely to contain objects, and that the object of interest covers a significant part of the interior of the image.

Another approach is to combine segmentation algorithms such as DeepMask [13] and Sharp-Mask [14] with an object detector shuch as MultiPathNet [22]. This results in object detections augmented with pixel level segmentations. From those segmentations rotated bounding boxes could be generated or, even better, distance measures can be made using the segmentations di-

**Figure 1.** *Example frames the recorded video.*

rectly. However this approach requires manual pixel level segmentations as training data while the suggested approach only requires a few manual clicks per training object.

## 1.2. Experimental setup

Video recordings were made using three Axis M3006-V cameras with a wide field of view, 134 degrees. They were placed in the ceiling at an height of 3.6-meters, pointing straight down to optimise overview over the study area. There is a significant overlap between the camera images to not miss events taking place at the border between the cameras. See Figure 1 for some example frames. In total 2315 hours (1 month) of 800x600 video at 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 result of this calibration is video images where the cows have the same size regardless of where in the image they appear. Also, the scan-lines 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

The classical pinhole camera model augmented with a lens distortion model was used to model the camera. The camera setup was calibrated by placing markers on the walls and stands in the middle of the room. They were all placed at the same height and thus defined a plane. The mean cow height in the barn was measured, and the plane was placed at their shoulder height. This height was estimated to be 1.49 meters with a standard deviation of 0.05 by measuring twelve random cows in the study area. This is the plane in which all of the landmarks considered below, except for the head, are expected to be found. By projecting detected landmarks back and forth between the camera images and this plane, detections from different cameras can be matched.

In addition to the markers, the focal length and lens distortion parameters were provided by the camera manufacturer. The lens distortion was removed, and a homograph was estimated that
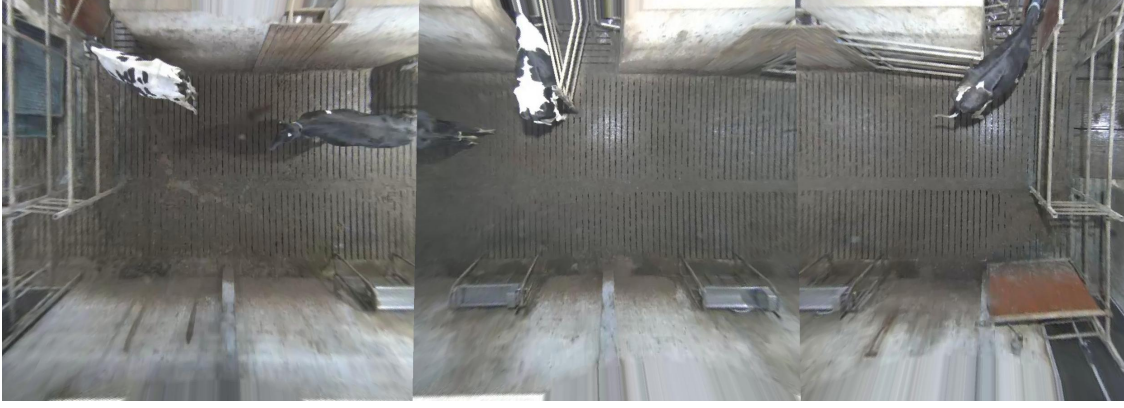
**Figure 2.** *The frames from Figure 1 projected onto the cow shoulder plane and stitched together*

projected each of the camera images onto the cow shoulder plane. Figure 2 shows a view stitched together from all three images in Figure 1. It shows an overview of the entire waiting area. At the borders between the cameras, the image becomes strange as cows there are viewed from different directions on opposite sides of the border. However, this image is only used for illustration. There is enough overlap between the images to allow them to be processed one by one and then the resulting detections can be combined using this calibration. Figure 3 shows the separate dewarped frames used by the detector. Note how the same cow is almost fully visible in both the left and the middle image.

By using homogeneous coordinates, the pinhole camera model that forms 2D image pixels, $x = (x_1, x_2, 1)$, by projecting world 3D points, $X = (X_1, X_2, X_3, 1)$, using a camera matrix, $P$, can be formed as

$$\lambda x = PX = K \begin{pmatrix} R & t \end{pmatrix} X = \begin{pmatrix} f & 0 & p_x \\ 0 & f & p_y \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} R & t \end{pmatrix} X, \tag{1}$$

were $f$ is the focal length of the camera and $(p_x, p_y)$ its principal point while $t$ and $R$ define its 3D position and orientation. This projective image is then distorted using some lens distortion function, $\hat{x} = f_{\text{dist}}(K^{-1}x)$. Here a radial fish eye lens distortion model is used. By using nominal image coordinates, $\lambda x_n = K^{-1}x$ and $\hat{x}_n = (\hat{x}_1 - p_x, \hat{x}_2 - p_y, 1)$ the inverse of this distortion



**Figure 3.** *Dewarped frames from each of the cameras with overlaps to allow the detector to process them one by one.*

function can be express as

$$|x_n| = \tan\left(\sum_i k_i |\hat{x}_n|^i\right), \tag{2}$$

where $k_i$ is the set of distortion parameters and $|x| = \sqrt{x_1^2 + x_2^2}$. Note that with this distortion model, the focal length becomes part of the distortion and Equation 2 allows the pixels produced by the camera $x$ to be transformed into nominal projective coordinates, $x_n$, directly. The distortion parameters, $k_i$, was provided by the manufacturer and the principal point, $(p_x, p_y)$, was assumed to be at the center of the image.

In the camera images, the markers placed in the cow shoulder plane was located manually and their nominal coordinates, $x_n$, were registered by clicking on them. Also, real world distances between the marks were estimated using a laser pointer. From these measurements the world coordinates, $x_s$, were calculated using multidimensional scaling [21]. One homograph, $H$, for each camera was fitted to the point correspondences allowing the pixels to be projected onto this plane using

$$\lambda x_d = H x_n. \tag{3}$$

This generates a common coordinate system for all of the cameras which allows detections from each of the cameras to be projected into this common frame. That way different cameras could be used in different parts of the waiting area. To minimise the amount of occlusion taking place, each part of the waiting area should use the closest camera to get a view from as straight above as possible. For that, the camera position in the cow shoulder plane needs to be estimated. That can be achieved by RQ factorisation of $H$ [3],

$$H = K_d R_d = \lambda \begin{pmatrix} f_x & s & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{pmatrix} R_d, \tag{4}$$

were the camera position is given by $c = (c_x, c_y, 1)$. Figure 2 shows an image where each pixel has been chosen from the camera closest to that pixel, i.e. with minimum $|x_d - c|$. The final crops used from each camera consists of the pixels in this stitched view with a border of $75$ pixels added on each side to make the overlap $150$ pixels, which roughly corresponds to the length of one cow.

## 3. Data annotation

The complete dataset contains three months of recordings collected from 3 cameras at 16 fps. That makes 400 million frames. Among these 1722 was randomly selected and annotated. The annotation process started before everything was recorded, so the video is sampled somewhat more densely in the beginning but the mean distance between two adjacent annotated frames is 30 min. So we believe that the annotated frames are fairly uncorrelated and should give a good representation of the distribution of frames that could be expected to be seen from these cameras.

The annotated 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.

| 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 + Softmax | 1x1 | 5 |

**Table 1** CNN architecture for landmark detection.

## 4.   CNN cow 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 [19] with batch normalisation [7] 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 $13 \times 13$ convolutional layer. Details are shown in Table 2.

The landmark net was trained on patches of $150 \times 150$ 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 $\pm 16$ 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 were initiated using random samples draw from a Gaussian distribution truncated at $2\sigma$, with standard deviation $\sigma = \sqrt{\frac{2}{n}}$, where $n$ is the number of inputs[4]. The networks were regularised with weight decay of $0.0001$ and optimised using stochastic gradient descent with $0.9$ momentum. The learning rate was initiated to $1.0$ and reduced by a factor

6

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

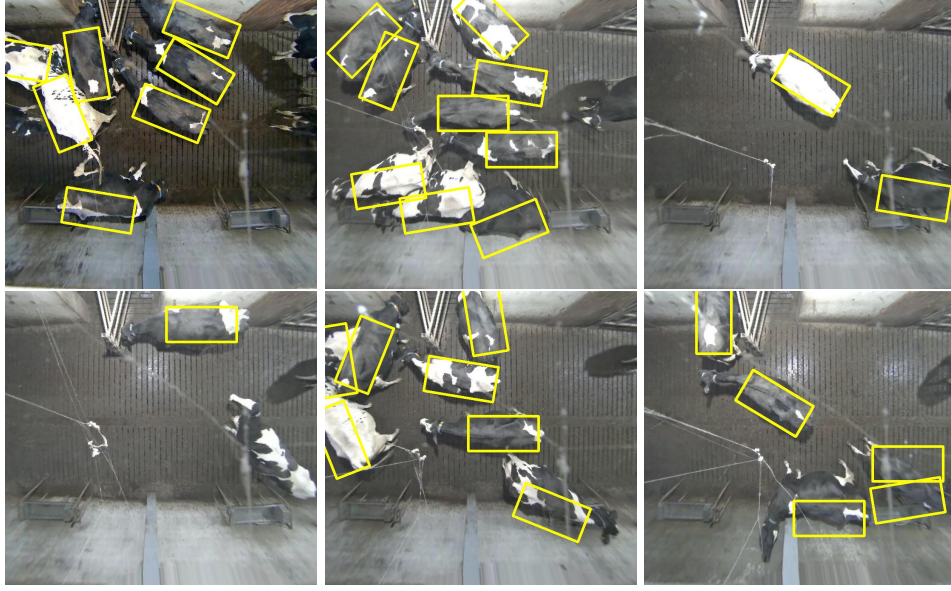**Table 2** CNN architecture to detect oriented cows



***Figure 4.*** *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).*

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 also to detect cows that are slightly outside the image.

Once the net was trained, the step length of the last maxpool layer was reduced from two to one 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

$$B(x, y, 32) = \begin{cases} 0 & \text{if} \quad \begin{aligned} \lfloor x_i \rfloor \leq x \leq \lceil x_i \rceil \\ \lfloor y_i \rfloor \leq y \leq \lceil y_i \rceil \end{aligned} \\ 1 & \text{otherwise} \end{cases} \tag{5}$$

7

**Figure 5.** *Example frames from a second recording made half a year after the first one.*

and $32$ orientation masks

$$B\left(x, y, c\right) = \begin{cases} 1 & \text{if} & \begin{array}{c} \lfloor x_i \rfloor - 1 \leq x \leq \lceil x_i \rceil + 1 \\ \lfloor y_i \rfloor - 1 \leq y \leq \lceil y_i \rceil + 1 \\ \text{adist}\left(\frac{2c\pi}{32}, c_i\right) < \frac{2\pi}{32} \end{array} \\ 0 & \text{otherwise} \end{cases} \quad (6)$$

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.

## 5. Watchdog evaluation

### 5.1. Cow detection

To evaluate the system, 6400 frames spread over the entire recording were processed by the CNN. A simple watchdog extracting frames containing two or more cows was implemented. That would be the most basic requirement for 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. A detection was considered correct if it's rotated bounding box overlapped more than 50% of the ground truth bounding box area.

Most, 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 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 false alarm rate of 0.9%. Some example detections are shown in Figure 4. 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.

### 5.2. Long term

Half a year after the original recordings were made another month of video was collected, c.f Figure 5, and the original detector was tested on this data. It did however not perform very well.

A set of 931 frames were sampled from the new recording and passed to the detector. It produced annotated images that were then inspected manually. Among these, 510 or 55% of the images were perfectly interpreted in the same sense as above, while 421 or 45% contained some sort of mistake. A lot of false detections were made. Especially in over exposed areas and there were also more misses.

Two main differences between this new data and the old one have been identified. First, the cameras have become quite dirty. They were cleaned a few times during the experiments, but they grow dirty quite fast, so for a system like this to become useful it will have to be able to handle somewhat dirty cameras. Second, the sun was a lot lower in the sky at the time of year of the new recording. That results in a different lighting situation. It is an indoor scene, but there are windows letting in the sun light.

From this new data, the 421 frames where the old detector made errors as well as 3 frames that contained no errors were manually annotated in the same way as before. They contained in total 2880 cows. The detector was retrained using both the new and the old data, which resulted in a detector with more even performance across the varying seasons. A validation set consisting of 10 % of the annotated data (196 frmaes) was separated out and not used during the training. This set was used to evaluate the new detector. In total it contained 408 cows, and 396 of these were detected with only 6 extra detections. That gives an cow hit rate of 97 % with a false alarm rate of 2.9 %. In terms of frames 92.8 % were perfectly interpreted while there were mistakes made in 14 of the frames. Among those 14, 13 contained more than one cow and thus the hit rate of the watchdog finding frames with two or more cows were 99.5 %.

## 5.3. Interaction detection

To remove even more of the uninteresting video, an additional feature, the minimum distance between the cows in the scene, was also extracted. Then a short sequence containing a lot of interesting interactions consisting of 187 frames evenly sampled over 10 minutes was extracted and manually annotated by an expert. Five different interactions were identified: body pushing, butting, head butting, head pressing and body sniffing. Frames where any of these interactions were present, were considered interesting and all other frames uninteresting. The minimum cow distance, $d_f$, for each frame $f$, was extracted. It is believed to be a good feature as cows need to be close to interact. Also, the cows needs to be close for some period of time, so we take the maximum over 9 consecutive frames and use as feature, $x_f = \max_{f-4 \leq i < f+4} d_i$ for frame $f$.

By thresholding $x_f$ a simple detector is formed. By choosing different thresholds, the results could be varied between not detecting any uninteresting frames and not missing any interesting frames. Figure 6 shows the amount of the interesting frames detected (true positives) as a function of the amount of the uninteresting frames remaining (false positives) for different thresholds. It is, for example, possible to discard 20% of the uninteresting frames while only losing 3% of the interesting once. If it is acceptable to lose more of the interesting frames, even more of uninteresting frames could be discarded as detailed by the curve.

Note that this is in addition to the filtering by a number of cows performed in Section 5.1. By combining the two filters, 50% of the uninteresting frames could be discarded while only loosing 4% of the interesting frames.
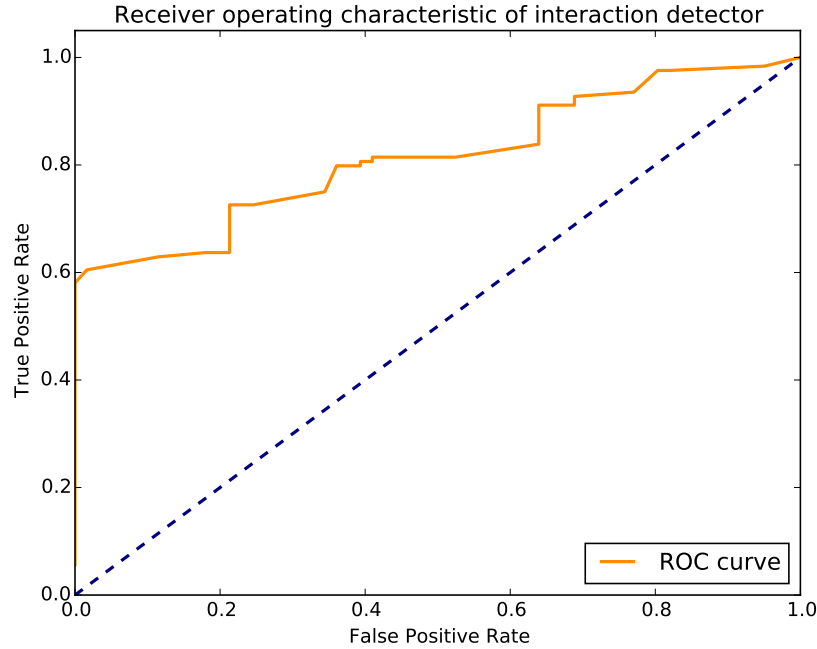
***Figure 6.*** *Results for different thresholds, with the amount of interesting frames kept (true positive rate) plotted as a function of the amount of uninteresting frames kept (false positive rate).*

## 6. 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 4. This detector was used to discard 38 % of the recorded video as uninteresting while only losing 1 % of the interesting video.

FIXME: season variantions "were highly affected by the shadows" =¿ "worst operating conditions for the CVBS occurred in the summer"

## 7. Acknowledgement

## 8. References

[1] Khalid Abdul Jabbar, Mark F Hansen, Melvyn Smith, and Lyndon Smith. Locomotion traits of dairy cows from overhead three-dimensional video. *Visual observation and analysis of Vertebrate And Insect Behavior*, 2016.

[2] 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.

[3] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, ISBN: 0521540518, second edition, 2004.

[4] 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.

[5] Shengfeng He and Rynson W.H. Lau. Oriented object proposals. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2015.

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

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

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

[9] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European Conference on Computer Vision*, pages 21–37. Springer, 2016.

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

[11] Carlos A Martinez-Ortiz, Richard M Everson, and Toby Mottram. Video tracking of dairy cows for assessing mobility scores. 2013.

[12] 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.

[13] Pedro H. O. Pinheiro, Ronan Collobert, and Piotr Dollár. Learning to segment object candidates. *CoRR*, abs/1506.06204, 2015.

[14] Pedro H. O. Pinheiro, Tsung-Yi Lin, Ronan Collobert, and Piotr Dollár. Learning to refine object segments. *CoRR*, abs/1603.08695, 2016.

[15] 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.

[16] Simona MC Porto, Claudia Arcidiacono, Umberto Anguzza, and Giovanni Cascone. The automatic detection of dairy cow feeding and standing behaviours in free-stall barns by a computer vision-based system. *Biosystems Engineering*, 133:46–55, 2015.

[17] Joseph Redmon, Santosh Kumar Divvala, Ross B. Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. *CoRR*, abs/1506.02640, 2015.

[18] Joseph Redmon and Ali Farhadi. Yolo9000: Better, faster, stronger. *arXiv preprint arXiv:1612.08242*, 2016.

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

[20] Porto S.M.C., Arcidiacono C., Anguzza U., and Cascone G. Detecting cows at the feed barrier by means of an image analysis algorithm. *International Conference RAGUSA SHWA 2012, September 3-6, 2012, Ragusa - Italy*.

[21] Gale Young and A. S. Householder. Discussion of a set of points in terms of their mutual distances. *Psychometrika*, 3(1):19–22, 1938.

[22] Sergey Zagoruyko, Adam Lerer, Tsung-Yi Lin, Pedro H. O. Pinheiro, Sam Gross, Soumith Chintala, and Piotr Dollár. A multipath network for object detection. *CoRR*, abs/1604.02135, 2016.