August 24, 2017

Dear Editor and Reviewers,

First, we would like to express our gratitude to the Editor and the Reviewers for taking their time to improve the paper. What follows is our attention to issues addressed.

> Reviewer: 1
>
> Comments to the Author The paper is well presented and describes an approach to discard uninteresting video footage for animal behaviour studies. The authors have presented this work previously in a workshop paper and have extended slightly their previous work. In particular the extensions covers the evaluation of the proposed technique on a second dataset captured in a different session; this constitutes a contribution of marginal significance. They have also described that previously omitted details on the camera calibration which it would be considered fundamental.
>
> The authors are annotating 7 landmarks points for each cow but only use 4 for their evaluation. The additional 3 landmarks could improve results and that could be considered a more significant contribution along with the evaluation over the complete dataset.

Approaches that would be able to utilize all 7 landmarks are indeed of interest. However extending the CNN used in this work to those particular 7 landmarks might be a bad idea as some of them are really close; The resolution of the landmark probability map is low due to all the max-pooling and close landmarks might compete for the same pixel there, which might harm performance. Also, for the particular watchdog criteria's studied in this paper, they would probably not provide much extra information anyway. So we argue that using all 7 landmarks would call for a different approach and would aid different goals which would make it a different paper.

Our point in this matter was not well described in the paper and we have added our reasoning to the paper (sec 4.1).

> More clarity is required on the dataset selection; the authors select 1722 frames that annotate manually for training purposes (Section 3), while they select 6400 frames to evaluate the system (Section 5.1). Are the two sets of frames mutually exclusive?

This is indeed a valid point. It was not ensured that none of the training frames were chosen as evaluation frams, but since they both were chosen randomly from a set of 400 million frames, the chance that they are mutually exclusive is 97.2%. Which, we argue, is sufficient to make our claims. This is now mentioned in the paper (sec 5.1).

> There classification output contains 32 different orientations plus one where a cow is not present. How many frames are in each of the 33 classes out of the 1722 frames?

Each frame can contain several cows. In total there are 6399 and all of them can be used as examples in each of the 32 classes thanks to the random rotations applied. We have added text to clarify this (sec 4.2).

> The authors state: "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." Not sure what this means and its significance; how does this affects the random sampling of the 1722 frames out of the 400 million frames.

We have added text to better clarify (sec 3).

> What constitutes a frame in terms of training/testing? Is it a single capture from 1 of the 3 cameras or the stitched scene (i.e. one that the complete containment area is in the field of view)? What is the input to the first stage CNN?

We have made several updates to Section 4 to address these questions.

> On section 5.1; it is claimed that 38% of recording is uninteresting; is this percentage computed from the 6400 frames? If so what is the percentage of frames that contained a single cow and consequently the percentage of frames with no cows? Maybe a histogram of number of cows per frame in the dataset would be useful to understand densities.

Great point. We have added such a histogram (Figure 5).

> The 100 frames that the watchdog selected (50) and discarded (50) is approximately 1.5% of the evaluation dataset. A higher proportion should be be evaluated.

Agreed. We have extended that 10 times to look at 1000 frames instead (sec 5.1).

> Considering that current practises by animal scientists, manually inspect the video footage, it would be more comprehensive and not onerous to evaluate across the collected data?

We agree that this would be desirable. However, we have in total a full year of recordings (4 months from each of 3 cameras). That is a bit more than what current practises typically look at and we simply do not have the resources in this project to look at it all unfortunately. Hence, we have to choose some sort of subset. We resorted to a random sampling across the entire dataset, which allows as much as possible of variability of the dataset to be evaluated.

> The percentage of discarded video footage (uninteresting footage) depends on the farming practises and mechanics of the particular farm. Therefore

> the 50% of discarded video footage is specific to that farm and that period of operation; this will vary from farm to farm, session, cow population and etc. This should be highlighted not to misguide the reader.

This is indeed a valid point. We have added text passages for clarify this (sec 6).

> Reviewer: 2
>
> Comments to the Author The abstract does not represent the content of the paper.

It has been updated to better represent the content.

> The comparison with earlier work using similar techniques is over emphasized given the limitations of the techniques.

Our aim here was to compare our work both with domain specific solution addressing similar problems and with more general techniques where the application at hand could be seen as a special case.

> Improvement of the English

We have improved and corrected English and grammar.

> Results No statistics or correlative analyses are given for any of the comparisons. summary of the experimental results is not explain properly.

We have made several changes to better describe and explain our evaluation, please see responses to reviewer 1 and 3.

> Reviewer: 3
>
> Comments to the Author It's unclear what the final output of the system actually is : The results are per-frame, but how many actual interactions are missed? Presumably, even if some frames are incorrectly classified, actual cow-cow interactions from other, adjacent, frames are still found ?

Good point, we've added an evaluation in terms of interactions as well (sec 5.3).

> The paper is generally well written. However there are a substantial number of typos (including double-quotes the wrong way around) and English grammar issues. I'd recommend getting native speaker to proof-read this thoroughly. Typos, grammar, quotes

We have improved and corrected English and grammar.

> You need to explain how the following is actually achieved : "The learning rate was initiated to 1.0 and reduced by a factor 10 each time the validation error flattens."

We've added such an explanation (sec 4.1).

> Some comments such as "A lot of false detections were made." need quantifying more thoroughly.

We've added an evaluation that quantifies this (sec 5.2)

> Citations: YOLO is perhaps more commonly cited as a CVPR paper, rather than the CoRR route ?

Good point. We have updated the reference.

Best regards

The Authors