A Cervix Detection Driven Deep Learning Approach for Cow Heat Analysis from Endoscopic Images

Ruiwen He, Halim Benhabiles, Feryal Windal, Gael Even, Christophe Audebert, Dominique Collard, Abdelmalik Taleb-Ahmed

A CERVIX DETECTION DRIVEN DEEP LEARNING APPROACH FOR COW HEAT ANALYSIS FROM ENDOSCOPIC IMAGES

Ruiwen He[⋆] *Halim Benhabiles*[⋆] *Feryal Windal*[⋆]
Gaël Even[†] *Christophe Audebert* [†] *Dominique Collard*[‡] *Abdelmalik Taleb-Ahmed*[⋆]

[⋆]Univ. Lille, CNRS, Centrale Lille, Junia, Univ. Polytechnique Hauts-de-France
UMR 8520 - IEMN - Institut d'Electronique de Microélectronique et de Nanotechnologie
F-59000 Lille, France
[†]GD Biotech - Gènes Diffusion, Lille 59000, France
[‡]LIMMS/CNRS-IIS The University of Tokyo, IRL 2820, Lille, France

ABSTRACT

In this article, we propose a new approach for the cow heat detection from endoscopic images. Our approach permits to identify on the fly the cow heat state through two successive stages, namely cervix detection then heat classification. For this purpose, images are analyzed by a Transformer based detection model to localize the cervix, in which case they are analyzed by a CNN-based heat classification model. The proposed approach permits to assist the farmer during the insemination operation by localizing the cervix in an accurate way. Moreover, the confidence level of the final decision of the classification model is increased by focusing its analysis only on cervix images. The effectiveness of our method is demonstrated on our generated dataset and the obtained performance outperform the state of the art.

Index Terms— Deep learning, Transformer, Endoscopic images, Artificial insemination, Cervix detection

1. INTRODUCTION

In cow farming, artificial insemination is a reproduction biotechnology that is widely spread [1]. Nevertheless, to successfully accomplish cow insemination, its heat period needs to be correctly detected, a stage that the farmer can possibly observe through certain behaviors of the cow [2], without complete certainty [3]. The expert or the veterinarian then intervenes to confirm the state of heat and locate the cervix to introduce the spermatozoa. At that point, farmers face two challenges, the availability of experts at the right time (cow heat) and the cost associated with their interventions, particularly if the state of heat is already over [4].
In this context, we have recently proposed the first vision system for cow heat detection based on the analysis of its genital tract [5]. More specifically, the system permits to analyze the endoscopic video, collected using an innovative insemination device named Eye Breed [6], and classify its heat state. The advantages of our system are two folds: 1) it offers the farmer the possibility to carry out autonomously the insemination operation thanks to the Eye breed device which is equipped with an endoscopic camera, 2) it provides the farmer assistance for identifying the cow heat state thanks to the video analysis model which is based on a tailored CNN classifier. In this article, we propose a new approach for cow heat detection. As illustrated in Figure 1, our approach permits to identify on the fly the cow heat state through two main stages namely cervix detection then heat classification. For this purpose, each frame of the input video stream, collected by the endoscopic camera of the Eye breed device, is analyzed by a Transformer based detection model to localize the cervix, in which case the frame is analyzed by the CNN based heat classification model. The final result of the frame analysis (cervix box + heat state) is displayed on the screen. Our proposed approach permits to significantly improve the two functionalities of our previous system [5]:

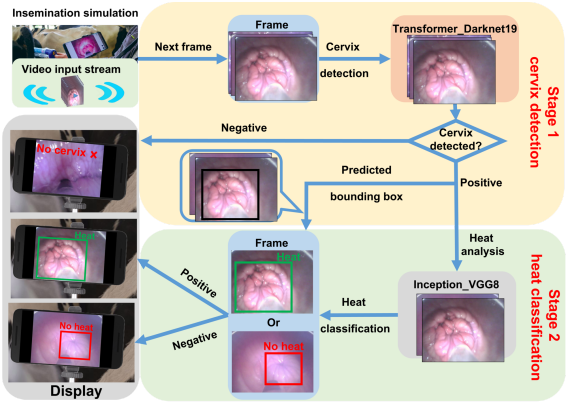

Fig. 1: Flowchart of our cow heat analysis method.

The authors gratefully acknowledge Claude Grenier, CEO Gènes Diffusion, Pierrick Drevillon CEO CECNA, Olivier Darasse, CEO Elexinn for the availability of data and the labeling of the videos.

- It further facilitates the insemination operation performed by the farmer thanks to the integration of the cervix detection model. Indeed, the model is able to detect and localize on the fly and in an accurate way the cow cervix offering thus to the farmer an assistance in the device guidance inside the cow genital tract.
- It increases the performance of our original heat state classification model by excluding from its analysis noisy video frames. To this end, the model only proceeds to the analysis of frames that are identified by the cervix detection model as positive.

Additionally, we show through the experimental study conducted on our dataset that the proposed Transformer architecture for object detection outperforms the state of the art ones [7, 8, 9, 10, 11, 12].

2. RELATED WORKS

Several works have been proposed in the literature for cow heat detection from the analysis of its visual behavior [13, 14, 15]. To this end, the proposed methods aim to exploit recorded videos in the farms in order to detect automatically special postures of the cows, such as being-mounted or standing-to-be-mounted, which are considered as the main heat indicators [3]. Basically, the proposed methods, exploit either handcrafted features to characterize video frames then analyze their spatio-temporal evolution to detect a behavior change [14, 15] or exploit CNN models trained for posture detection [13]. However, the exploitation of these methods in the context of farms represents a real challenge in reason of cameras deployment issues both on ethical and logistic plans. Moreover, as the cow behavior is impacted by its health and its environment [3], the effectiveness of these methods remains limited. Our proposed method permits to address these issues since it is based on the analysis of the genital tract of the cow. The main novelty of the present work compared to our previous work [5] is the proposition of a two stages methodology for heat analysis namely 1) cervix detection and 2) heat classification instead of a one-stage methodology based on the classification only. The new methodology permits to increase the heat analysis precision. For this purpose, we designed a Transformer based cervix detection model. In what follows, we will briefly discuss existing works on object detection and the difference with our detection model.

As raised in the survey made by Liu et al. [16], most of the recent methods are based on a unified detection strategy where CNN architectures are exploited to simultaneously predict object bounding boxes and their associated classes. In this sense, among the existing CNN architectures one can cite the YOLO series 1-5 [17, 11, 10, 9, 7], the transformer-based DETR model [8] and Deformable-DETR [12]. Indeed, these methods have shown high performances in term of detection accuracy over the benchmarking datasets such as the COCO [18] and the VOC [19] datasets. Nevertheless, in reason of the requirements of the existing datasets [18, 19], these architectures are designed and trained in such a way to detect at least one target object in the image. Hence, their direct exploitation for cervix detection from endoscopic videos will lead to a high detection rate of false positives which is a major issue in our case. Indeed, detecting a false cervix will increase the number of noisy frames to be analyzed by the classifier and will mislead the farmer in the insemination operation. To address this issue, our detection model has been designed to perform exclusively the detection task and has been trained on a balanced image dataset that includes positive and negative cervix examples.

3. PROPOSED METHODOLOGY

Our proposed methodology for cow heat analysis exploits two CNN architectures namely Transformer-Darknet19 and Inception-VGG8 for cervix detection (stage 1) and heat classification (stage 2) respectively. In this section, we present the first architecture (Transformer-Darknet) which is a variant of the original DETR architecture [8]. For more details about the second architecture (Inception-VGG8), we refer the reader to our recent work [5]. Figure 2 illustrates the design of

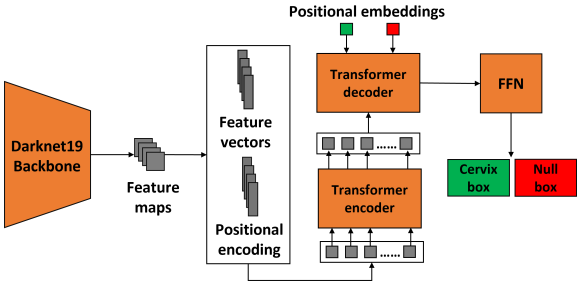

Fig. 2: Overview of our Transformer-Darknet19 architecture for cervix detection.

our Transformer-Darknet19 architecture. More specifically, image features are extracted using a CNN backbone corresponding to the Darknet19 [11] architecture. The features are then flattened and given to the encoder of the transformer in the form of a data sequence which is the expected form by transformers. To avoid losing spatial positioning information of the image pixels, positional encodings of the flattened features are joined to the input data sequence of the encoder. The encoder proceeds then to learn, through a self-attention mechanism [3], how to focus on relevant object patterns from the input sequence. The encoder output is then exploited by the decoder to learn, through an attention mechanism, how it can affect the prediction of the two positional embeddings that represent the existence and absence of a cervix object respectively. Finally, the decoded output is passed to an FFN (Feed Forward network corresponding to a 3-layer perceptron

of size 2048 each) for box prediction.
It is worth mentioning that to adapt the original DETR architecture [8] to our problem, we have: 1) used the Darknet CNN backbone for image features extraction instead of ResNet, 2) limited the number of encoders and decoders to one instance instead of six, 3) limited the number of positional embeddings for encoding the cervix existence and absence to 2 instead of 100.
The choice of Darknet19 backbone is explained by the fact that, contrary to ResNet, it has a reasonable depth which helps to limit the overfitting during the training process of the whole architecture. The backbone has been firstly pretrained on the ImageNet dataset and to extract image features we considered the output of the 18th convolutional layer (1024 feature maps). The limitation of the number of encoders/decoders to 1 is due to the particularity of our detection problem. Indeed, in our case, endoscopic images contain at most one target object of a unique class (cervix). Hence one encoder/decoder is enough to characterize the object while reducing the complexity of the global architecture. To achieve our goal of encoding the existence and absence of the cervix (2 positional embeddings), we have trained our global architecture on a balanced dataset that includes images of the two classes (with and without cervix boxes). In addition, we have modified the loss function by replacing the L_1 loss component by the $smooth_{L1}$ loss in order to faster the convergence of the learning [20].

4. EXPERIMENTAL STUDY

4.1. Generated endoscopic image dataset

Our dataset consists of 12732 labeled endoscopic images which have been extracted from 79 recorded videos of simulated insemination operation on several cows using the Eye breed device. The labelling of each image corresponds to i) its heat class (heat or no-heat) which is the video class assigned by the expert and ii) associated cervix bounding box which is set up to null if there is no cervix. The set of images have been split into a training set of 10734 images and a validation set of 1998 images. Both two sets are balanced in term of positive and negative cervix boxes. In addition, the dataset split has been done in such a way that the frames of a given video are completely included either in the training set or in the validation set. To reach this goal, we have randomly split our set of videos into 69 videos for training and 10 for validation. All the video frames have been extracted using an FPS rate set to 5. To evaluate the generalization level of our models, we exploited the 10 validation videos to extract a larger set of labelled images namely 32552 and considered it as a test set. For this purpose, the video frames have been extracted using the default FPS rate of the concerned videos which is ranged between 20 and 102. Indeed, this parameter can be finetuned by the operator during the recording.

4.2. Performance evaluation

To evaluate the performance of our method we considered two scenarios: 1) we have evaluated the performance of the cervix detection model (stage 1) and 2) we have evaluated the performance of the global pipeline namely cervix detection and heat classification (stage 1 + stage 2). The obtained performances in both scenarios have been compared with those of the state of the art namely YOLO series [7, 9, 10, 11], DETR-ResNet50/101 [8], Deformable-DETR-R50 [12] and Inception-VGG8 [5]. To this end, each method has been trained, validated and tested on the sets presented in the previous section. To train the state-of-the-art methods we used their respective source codes which are publicly available and set up their parameters following the recommendations from the referenced articles.

4.2.1. Cervix detection performance

Positive vs. Negative frames – To measure the ability of the models to distinguish between positive and negative cervix frames we calculated their accuracy ($ACC = (TP+TN)/(TP+TN+FN+FP)$), precision ($PRE = TP/(TP+FP)$) and recall ($REC = TP/(TP+FN)$). Table 1 summarizes the obtained results on the 32552 images of the test set. We can observe that our detection model reached an accuracy of 87.1% which is the best one compared to the other models. The table also shows that, contrary to the other models, our model tends to favor the precision over the recall (98.9% vs. 76.3%). This means that our model has a weak rate of FP which is more suited in the context of our insemination application.
Cervix box localization – To measure the ability of the models to detect the cervix and localize it, we calculated their average precision ($AP = TP/(TP+FP+FN)$). The TP, FP and FN are calculated with respect to the ground-truth boxes based on an IoU (Intersection over Union) metric. They correspond for a given frame to what follow:

- TP a box detected with an IoU$\geq$ threshhold compared to the ground-truth box.
- FP all additional boxes detected within the same frame.
- FN missed box (empty frame) or box detected with an IoU$<$threshold compared to the ground-truth box.

Figure 3 shows the obtained results of the top 3 models on the 16773 positive frames of the test set using several IoU thresholds [0.5, 0.95]. The curves in the figure permit to observe that our model reached a precision which is slightly less than the one obtained by the YOLOv5s (75.6 vs. 79.5) for a small IOU threshold (0.5). Nevertheless, our model and contrary to the others, succeeded to keep the same level of performance with relatively a high threshold (0.75) which clearly indicates that it has a better cervix localization ability.

	YOLOv2 [11]	YOLOv3 [10]	YOLOv4 [9]	YOLOv5s [7]	DETR-R50 [8]	DETR-R101 [8]	D-DETR-R50 [12]	Our method
ACC	51.4	62.5	78.1	83.8	81	51.8	67.3	**87.1**
PRE	51.4	67.7	78.1	78.9	80.4	58.2	63.7	**98.9**
REC	**100**	95.8	85.1	93.5	83.4	23.3	83.2	76.3

Table 1: Cervix detection models performance comparison obtained on the test set (32552 images).

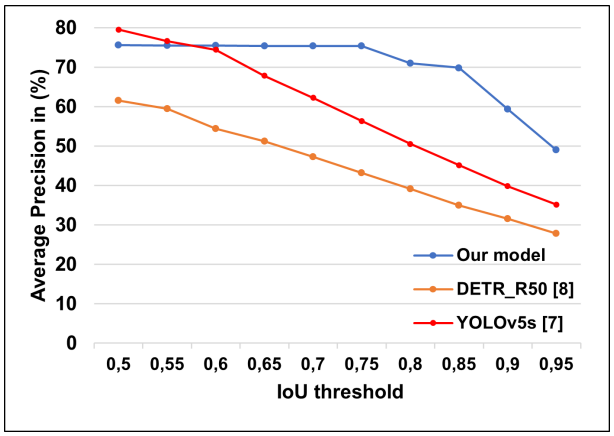

Fig. 3: Performance comparison in term of cow cervix localization obtained on the test set of positive frames (16773 images).

4.2.2. Cervix detection and heat classification performance

We measured for each model the mean of average precision ($MAP = (AP_{heat} + AP_{no-heat})/2$) which takes into consideration the detection quality and classification as well. Indeed, it is not relevant to evaluate separately the classification of the state of the art methods in reason of their unified detection and classification strategy.
The $AP_{class} = TP_{class}/(TP_{class} + FP_{class} + FN_{class}))$ where $class = heat, no\text{-}heat$. To this end, we used the 16773 positive frames of the test set and split them according to their heat category. Each AP_{class} has been calculated using several IoU thresholds [0.5,0.95] and has been averaged to obtain a global performance for each class. Table 2 shows the obtained results by the best 3 models together with their complexities. We can observe that our method has reached the highest percentage for both classes outperforming widely the state of the art ones. In addition, it has a reasonable complexity making possible its deployment on a smartphone. To compare the performance of our new method with our previous one [5], we evaluated both of them on the validation set of 10 videos and calculated their rates of correct prediction (heat or no-heat). We also exprimented two other conbinations : DETR-R50 [8] + IVGG8 [5] and YOLOv5s [7] + IVGG8 [5]. Figure 4 shows the obtained results on 5 videos from each class. We can observe that all methods are able to correctly predict the heat state. However, our new method shows more confidence in its decision since it gives the highest rates for all the videos.

Method	AP_h	AP_{nh}	MAP	Params	FPS
YOLOv5s [7]	59.6	56.2	57.9	**7M**	**151**
DETR-R50 [8]	44.5	40.6	42.5	36.7M	20
Our method	**68.5**	**70.7**	**69.6**	26.5M	46

Table 2: Performance comparison of the final decision on cervix detection and heat classification obtained on the test set of positive frames (16773 images).

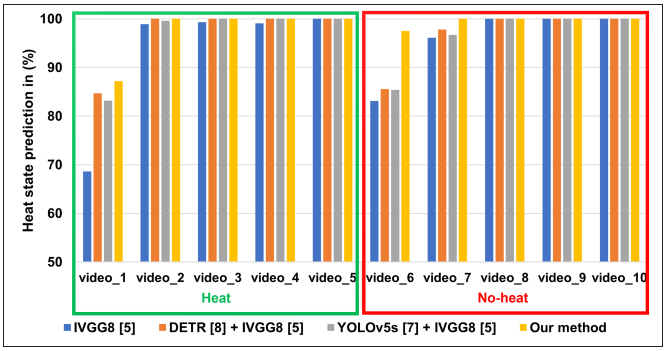

Fig. 4: Results in (%) of heat state prediction on the validation set composed of 10 videos.

5. CONCLUSION

A new deep learning-based approach for cow heat analysis from endoscopic images has been proposed. The approach goes through two main stages namely cervix detection and heat classification. The effectiveness of our approach has been demonstrated on our dataset outperforming the state-of-the-art methods. More specifically, our transformer-based detection model reached an accuracy of 87.1% which permitted to increase the confidence level of the final decision of our method for heat prediction in comparison with our previous method [5]. We believe that this new method will further assist the farmer in the insemination operation offering him a precision detection and analysis tool.

Funding

This project has been funded by the FEDER European program, JUNIA French Engineering school and Gènes Diffusion French company.

References

[1] DP Berry, SC Ring, AJ Twomey, and RD Evans, "Choice of artificial insemination beef bulls used to mate with female dairy cattle," *Journal of dairy science*, vol. 103, no. 2, pp. 1701–1710, 2020.

[2] Ina Gaude, Andreas Kempf, Klaas Dietrich Strüve, and Martina Hoedemaker, "Estrus signs in holstein friesian dairy cows and their reliability for ovulation detection in the context of visual estrus detection," *Livestock Science*, vol. 245, pp. 104449, 2021.

[3] Hawar M Zebari, S Mark Rutter, and Emma CL Bleach, "Characterizing changes in activity and feeding behaviour of lactating dairy cows during behavioural and silent oestrus," *Applied Animal Behaviour Science*, vol. 206, pp. 12–17, 2018.

[4] Teweldemedhn Mekonnen and Leul Berhe, "Assessment on artificial insemination service delivery system, challenges and opportunities of artificial insemination services in cattle production in western zone of tigray region, ethiopia," *International Journal of Livestock Production*, vol. 11, no. 4, pp. 135–145, 2020.

[5] Ruiwen He, Halim Benhabiles, Feryal Windal, Gaël Even, Christophe Audebert, Agathe Decherf, Dominique Collard, and Abdelmalik Taleb-Ahmed, "A cnn-based methodology for cow heat analysis from endoscopic images," *Applied Intelligence*, pp. 1–14, 2021.

[6] Agathe Decherf and Pierrick Drevillon, "Device for the atraumatic transfer of a material or substance with a reproductive, therapeutic or diagnostic purpose into female mammals," June 9 2020, US Patent 10,675,133.

[7] Glenn Jocher, "ultralytics/yolov5: v3.1 - Bug Fixes and Performance Improvements," Oct. 2020, accessed on 18 May 2020.

[8] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko, "End-to-end object detection with transformers," in *European conference on computer vision*. Springer, 2020, pp. 213–229.

[9] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao, "Yolov4: Optimal speed and accuracy of object detection," *arXiv preprint arXiv:2004.10934*, 2020.

[10] Joseph Redmon and Ali Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[11] Joseph Redmon and Ali Farhadi, "Yolo9000: better, faster, stronger," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7263–7271.

[12] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai, "Deformable detr: Deformable transformers for end-to-end object detection," in *International Conference on Learning Representations*, 2020.

[13] Jung-woo Chae and Hyun-chong Cho, "Identifying the mating posture of cattle using deep learning-based object detection with networks of various settings," *Journal of Electrical Engineering & Technology*, vol. 16, no. 3, pp. 1685–1692, 2021.

[14] Yangyang Guo, Ziru Zhang, Dongjian He, Jinyu Niu, and Yi Tan, "Detection of cow mounting behavior using region geometry and optical flow characteristics," *Computers and Electronics in Agriculture*, vol. 163, pp. 104828, 2019.

[15] Shogo Higaki, Kei Horihata, Chie Suzuki, Reina Sakurai, Tomoko Suda, and Koji Yoshioka, "Estrus detection using background image subtraction technique in tie-stalled cows," *Animals*, vol. 11, no. 6, pp. 1795, 2021.

[16] Li Liu, Wanli Ouyang, Xiaogang Wang, Paul Fieguth, Jie Chen, Xinwang Liu, and Matti Pietikäinen, "Deep learning for generic object detection: A survey," *International journal of computer vision*, vol. 128, no. 2, pp. 261–318, 2020.

[17] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[18] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.

[19] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman, "The pascal visual object classes (voc) challenge," *International journal of computer vision*, vol. 88, no. 2, pp. 303–338, 2010.

[20] Ross Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.