Towards Self-Supervision for Video Identification of Individual Holstein-Friesian Cattle: The Cows2021 Dataset

Jing Gao1 Tilo Burghardt1 William Andrew1,2 Andrew W. Dowsey2 Neill W. Campbell1

1Department of Computer Science University of Bristol Bristol, United Kingdom
2Bristol Veterinary School University of Bristol Bristol, United Kingdom

Abstract

In this paper we publish the largest identity-annotated Holstein-Friesian cattle dataset (Cows2021) and a first self-supervision framework for video identification of individual animals. The dataset contains 10,402 RGB images with labels for localisation and identity as well as 301 videos from the same herd. The data shows top-down in-barn imagery, which captures the breed’s individually distinctive black and white coat pattern. Motivated by the labelling burden involved in constructing visual cattle identification systems, we propose exploiting the temporal coat pattern appearance across videos as a self-supervision signal for animal identity learning. Using an individual-agnostic cattle detector that yields oriented bounding-boxes, rotation-normalised tracklets of individuals are formed via tracking-by-detection and enriched via augmentations. This produces a ‘positive’ sample set per tracklet, which is paired against a ‘negative’ set sampled from random cattle of other videos. Frame-triplet contrastive learning is then employed to construct a metric latent space. The fitting of a Gaussian Mixture Model to this space yields a cattle identity classifier. Results show an accuracy of Top-1: 57.0% and Top-4: 76.9% and an Adjusted Rand Index: 0.53 compared to the ground truth. Whilst supervised training surpasses this benchmark by a large margin, we conclude that self-supervision can nevertheless play a highly effective role in speeding up labelling efforts when initially constructing supervision information. We provide all data and full source code alongside an analysis and evaluation of the system.

1. Introduction and Background

Holstein-Friesians are, with a global population of 70 million [16] animals, the most numerous and also highest milk-yielding [41] cattle breed in the world. Cattle identification (ID) via tags [19, 12, 39] is mandatory [32, 31], yet transponders [23], branding [1, 9] and biometric ID [25] via face [11], muzzle [34, 26, 22, 42, 10, 14], retina [2], rear [36], or coat patterns [29, 20, 27, 8] are also viable. The last can conveniently operate from a distance above and has recently been implemented via supervised deep learning [6, 3]. However, research into reducing manual labelling efforts for creating and maintaining such ID systems is in its infancy [5, 44]. Particularly, unsupervised learning for coat pattern identification of Holstein-Friesians has not been tried and public datasets [4] are small to date.

This paper addresses these shortcomings and introduces the largest ID-annotated dataset of Holstein-Friesians: Cows2021 so far, alongside a basic self-supervision system for video identification of individual animals (see Fig. 1).
2. Dataset Cows2021

We introduce the RGB image dataset Cows2021\(^1\), which features a herd of 186 Holstein-Friesian cattle (see Fig. 2) and was acquired via an Intel D435 at University of Bristol’s Wyndhurst Farm in Langford Village, UK. The camera pointed downwards from 4m above the ground over a walkway (see Fig. 3) between milking parlour and holding pens. Motion-triggered recordings took place after milking across 1 month of filming.

The dataset is resolved at 1280 × 720 pixels per frame with 8bit per RGB channel. It contains 10, 402 still images, in addition to 301 videos (each of length 5.5s) at 30fps. The distribution of stills across individuals and time reflects the natural workings of the farm (see Fig. 4). Various expert ground truth (GT) annotations are provided alongside the acquired dataset.

**Oriented Bounding-Box Cattle Annotations.** Adhering to the VOC 2012 guidelines [15] for object annotation, we manually labelled\(^2\) all visible cattle torso instances. Significantly clipped torso instances were (following [15]) not used further and given a ‘clipped’ tag. Example images from the resulting set of 13,165 non-clipped cattle torso annotations are given in red in Fig. 3 (bottom). Each oriented bounding-box label is parameterised by a tuple: \((c_x, c_y, w, h, \theta)\) corresponding to the box centrepoint, width, height, and head direction.

**Animal Identity Annotations.** Overall 13,784 detected (see Sec. 3) cattle instances were manually ID-assigned to one of 182 individuals (see Fig. 2). The 4 all-black cows were excluded from the ID study subject to future research. The number of occurrences of individuals varies from 2 to 273 with mean \(\mu = 77.5\) and standard deviation \(\sigma = 39.9\) (see Fig. 4). 8,670 of these annotations were filmed on different days to the video data. These were used to form the identity test data.

**Video Data and Tracklet Annotations.** In addition to still images, the dataset contains videos with tracklet information designed for utilisation as a rich source of self-supervision in identity learning. Using a highly reliable ID-agnostic cattle detector (see Sec. 3) and sampling at 5Hz, tracking-by-detection was employed to connect nearest centrepoints of detections in neighbouring frames and thereby extract entire tracklets of the same individual (see Fig. 1). Manual checking ensured no tracking errors occurred. The average number of tracklets per video is 1.45.
have a fully-connected final layer to learn a latent ResNet50 [17] pretrained on ImageNet [13], modified to

4.1. Contrastive Training

consuming task of manual labelling.

the learning of filmed individual animals to aid the time-

investigated how far this data can be used to self-supervise

available in-barn videos of a Holstein-Friesian herd. We in-

able tracklets can be generated (see Sec. 2) from readily

translating in-barn videos to tracklets.

4.2. Animal Identity Discovery via Clustering

Existing multi-object single-frame cattle detectors [11, 7, 5] produce image-aligned bounding-boxes that cannot

avoid capturing several individuals in crowded scenes (see Fig. 3), which is problematic for subsequent identity assignment. In response, we constructed a first orientation-aware cattle detector (see Fig. 3 blue) by modifying RetinaNet [28] with an ImageNet-pretrained [13] ResNet50 backbone [17]. We added additional target parameters for orientation encoding and rotated anchors implemented in 5 layers (P3 - P7). To train the network, we partitioned the still image set approximately 7 : 1 : 2 for training, validation and testing, respectively. We used timestamps to split data so any temporal bias is reduced. We then trained the network against Focal Loss [28] with settings \(\gamma = 2, \alpha = 0.25, \lambda = 1\) via SGD [38] with a learning rate of \(1 \times 10^{-5}\), momentum of 0.9 [35], and weight decay of \(1 \times 10^{-4}\). Fig. 5 illustrates training and depicts full performance benchmarks for the detector. For the test set, it operates at an Average Precision of 97.3% using an Intersection over Union (IoU) threshold of 0.7, reliably translating in-barn videos to tracklets.

4. Self-Supervised Animal Identity Learning

Given an ID-agnostic cattle detector (see Sec. 3), reliable tracklets can be generated (see Sec. 2) from readily available in-barn videos of a Holstein-Friesian herd. We investigated how far this data can be used to self-supervise the learning of filmed individual animals to aid the time-consuming task of manual labelling.

4.1. Contrastive Training

Identification Network and Triplet Loss. We use a ResNet50 [17] pretrained on ImageNet [13], modified to have a fully-connected final layer to learn a latent 128-dimensional ID-space. Across the training data of all videos, we normalise each tracklet for rotation (as seen in Fig. 2) and organise it into a ‘positive’ ID sample set representing the same, unknown individual. We pair this set against ‘negative’ samples from random cattle of other videos, which have a high chance of containing a different individual. All sets are enhanced via rotational augmentation (max. angle \(\pm 7^\circ\)). The separate image data was used as a validation and testing base, split 1 : 3. Reciprocal triplet loss (RTL) [30] is then employed for learning an ID-encoding latent space via an online batch hard mining strategy [18]:

\[
\mathbb{L}_{RTL} = d(x_a, x_p) + \frac{1}{d(x_a, x_n)}
\]

where \(x_a\) and \(x_p\) are sampled from the ‘positive’ set and \(x_n\) is a ‘negative’ sample. We trained the network for 7 hours via SGD [38] over 50 epochs with batch size 16, learning rate \(1 \times 10^{-4}\), margin \(\alpha = 2\), and weight decay \(1 \times 10^{-4}\). The pocket algorithm [40] against the validation set was used to tackle overfitting (see Fig. 6).

4.2. Animal Identity Discovery via Clustering

Clustering. We then fitted [33] a Gaussian Mixture Model (GMM) [37] to the generated 128-dimensional space by setting the cluster cardinality to the known \(k = 182\) patterned individual animals with 200 iterations. Resulting clusters are then interpreted as representing separate animal identities. A t-distributed Stochastic Neighbour Embedding (t-SNE) [43] of the training set projected into the clustered space is visualised in Fig. 7. In order to evaluate the clustering performance, we used two measures: the Adjusted Rand Index (ARI) [21] and ID prediction accuracy. For the latter, each GMM cluster is assigned to the one individual ID with the highest overlap which is defined as:

\[
\mathcal{O}_t = C / L
\]

where \(C\) is the number of images in a GMM cluster that belong to an individual, and \(L\) is the total number of images of the individual. This produces (GMM Cluster)-(ID Label) pairs for accuracy evaluation.

Top-N Accuracy. In order to quantitatively evaluate the capacity to aid human annotation, we consider a scenario
where a user annotates IDs as a one-out-of-N pick (expanding N if the correct ID is not present). Thus, the Top-N system accuracy [24] is a key measure to investigate. For each cluster one can rank all identities according to $O_l$. Identities that have a $O_l = 0$ form the randomly assigned tail of the sequence. For every data point this provides a general Top-N assigned ID. Finding the GT identity amongst the Top-N assigned IDs is then counted as correct identification.

5. Experimental Results and Discussion

Structural Clustering Similarity. In order to characterise the ID performance as if this were a new, unknown herd, we calculated the ARI to be 0.53 for the test set when measured between the partitioning derived from the clustering provided by the GMM versus the identity GT. This measure captures the (purely structural) similarities between the two clusterings.

Clustering Accuracy. In order to characterise the ID performance with class labels, we calculated Top-N accuracy for the test set as depicted in Table 1. Figure 8 visualises the identification performance and misclassification severity using a t-SNE plot.

Context and Result Discussion. Considering that 182 classes were used and absolutely no training labelling was provided, results of 57.0% Top-1 accuracy and 76.9% Top-4 accuracy are an encouraging and practically relevant first step towards self-supervision in this domain. We now that individual Holstein-Friesian identification via supervised deep learning is a widely solved task with systems achieving near-perfect benchmarks when using multi-frame LRCNs [7] and good results even in partial annotation settings [5]. However, labelling efforts are laborious for supervised systems of larger herds; they require days if not weeks of manual annotation effort using visual dictionaries of animal ground truth. Humans can efficiently compare small sets of images. Thus, using the described pipeline we could present the user with a set of (e.g. 4) images that contain the correct individual with a chance better than 3-in-4. As part of a toolchain, the approach presented can potentially dramatically reduce labelling times and help bootstrap production systems via combinations of self-supervised learning followed by open set fine-tuning [5].

<table>
<thead>
<tr>
<th></th>
<th>N=1</th>
<th>N=2</th>
<th>N=4</th>
<th>N=8</th>
<th>N=16</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy (%)</td>
<td>57.0</td>
<td>71.8</td>
<td>76.9</td>
<td>79.7</td>
<td>81.8</td>
</tr>
</tbody>
</table>

Table 1: Top – N Performance. Shown is ID accuracy for a variety of N across all 8,670 test instances of the 182 identities.

6. Conclusion

In this paper we presented the largest identity-annotated Holstein-Friesian cattle dataset, Cows2021, made available to date. We also showed a first self-supervision framework for identifying individual animals. Driven by the enormous labelling effort involved in constructing visual cattle identification systems, we proposed exploiting coat pattern appearance across videos as a self-supervision signal. A generic cattle detector yielded oriented bounding-boxes which were normalised and augmented. Triplet loss contrastive learning was then used to construct a latent space wherein we fitted a GMM. This yielded a cattle identity classifier which we evaluated. Our results showed that the achieved accuracy levels are strong enough to help speed up ID labelling efforts for supervised systems in the future. Despite the need for even larger datasets, we hope that the published dataset, code, and benchmark will stimulate research in the area of self-supervision learning for biometric animal (re)identification.

Acknowledgements. This work was supported by The Alan Turing Institute under the EPSRC grant EP/N510129/1 and the John Oldacre Foundation through the John Oldacre Centre for Sustainability and Welfare in Dairy Production, Bristol Veterinary School. Jing Gao was supported by the China Scholarship Council. We thank Kate Robinson and the Wyndhurst Farm staff for their assistance with data collection.
References


