
Peer reviewed version

Link to published version (if available): 10.1109/WACVW50321.2020.9096949

Link to publication record in Explore Bristol Research
PDF-document

This is the author accepted manuscript (AAM). The final published version (version of record) is available online via Institute of Electrical and Electronics Engineers at https://ieeexplore.ieee.org/document/9096949 . Please refer to any applicable terms of use of the publisher.

University of Bristol - Explore Bristol Research

General rights

This document is made available in accordance with publisher policies. Please cite only the published version using the reference above. Full terms of use are available: http://www.bristol.ac.uk/red/research-policy/pure/user-guides/ebr-terms/
Fusing Animal Biometrics with Autonomous Robotics:
Drone-based Search and Individual ID of Friesian Cattle (Extended Abstract)

William Andrew Colin Greatwood Tilo Burghardt
University of Bristol
{will.andrew, colin.greatwood, tb2935}@bristol.ac.uk

Abstract

This work covers the robotic drone integration of a re-identification system for Friesian Cattle. We have built a computationally-enhanced M100 UAV platform with an onboard deep learning inference system for integrated computer vision and navigation able to autonomously find and visually identify by coat pattern individual Holstein Friesian cattle in freely moving herds. For autonomous drone-based identification we describe an approach that utilises three deep convolutional neural network architectures running live onboard the aircraft; that is, a YoloV2-based species detector, a dual-stream Convolutional Neural Network (CNN) delivering exploratory agency and an InceptionV3-based biometric Long-term Recurrent Convolutional Network (LRCN) for individual animal identification. We evaluate the performance of components offline, and also online via real-world field tests of autonomous low-altitude flight in a farm environment. The presented proof-of-concept system is a successful step towards autonomous biometric identification of individual animals from the air in open pasture environments and inside farms for tag-less AI support in farming and ecology. The work is published in full in IROS 2019 [4].

1. Introduction

This paper presents an unmanned aerial vehicle (UAV) platform with onboard deep learning inference (see Fig. 1) that autonomously locates and visually identifies individual Holstein Friesian cattle by their uniquely coloured coats in low altitude flight (approx. 10m) within a geo-fenced farm area. The task encompasses the integrated performance of species detection, exploratory agency, and individual animal identification (ID). All tasks are performed entirely onboard a custom DJI M100 quadrotor with limited computational resources, battery lifetime, and payload size. In doing so, this work attempts to assist agricultural monitoring in performing minimally-invasive cattle localisation and identification in the field. Possible applications include the behavioural analysis of social hierarchies [46, 19, 35] grazing patterns [13, 14] and herd welfare [43]. The search for targets with unknown locations traditionally arises in search and rescue (SAR) scenarios [25, 47, 40]. In visually-supported navigation for this task, approaches broadly operate either a map-based or map-less paradigm [6, 9]. Map-less approaches have no global environment representation and traditionally operate using template appearance matching [29, 18], optical-flow guidance [42], or landmark feature tracking [42, 41]. More recently, such systems have been replaced with visual input classification via CNNs [12, 37]. In this work, we build on a simulation setup presented in [3] and formulate a 2D global grid approximation of the environment (see map M in Fig. 1) for storing visited positions, current location, and successful target recoveries. This concept is inspired by occupancy grid maps [7, 34], as opposed to post-exploration maps [32] or topological maps [20]. For our cattle recovery task – and despite their simplicity – grid maps still represent a highly effective tool [3] for exploring the solution space of AI solutions [49, 30]. Coat pattern identification of individual Friesian cattle represents a form of animal biometrics [23]. Early systems for the particular task at hand utilised the Scale-Invariant Feature Transform (SIFT) [26] on image sequences [27] or Affine SIFT (ASIFT) [33] to map from dorsal cow patterns [5] to animal IDs [35]. However, for up-to-date performance we base our individual ID component on recent CNN-grounded biometric work [2] where temporal stacks of the regions of interest (RoIs) around detected cattle are analysed by a Long-term Recurrent Convolutional Network (LRCN) [10] as shown in Figure 1 in red. This architecture represents a compromise between light-weight onboard operation and more high-end networks with heavier computational footprints. Whilst aerial wildlife census applications routinely use manually controlled UAVs [16, 21, 1, 39] and have experimented with widely automated photographic gliders [31]. To the best of our knowledge this line of work presents the first proof-of-concept system for fully autonomous exploration and online individual biometric identification of animals completely onboard an aircraft.

Acknowledgements: the workshop presentation of this extended abstract was made possible by funding from the Artificial Intelligence grant from the Alan Turing Institute led by Professor Andrew Dowsey.
For individual Friesian cattle search and identification we propose an approach that utilises three deep convolutional neural architectures operating onboard a computationally-enhanced DJI Matrice 100 platform to achieve species detection (green), exploratory agency (blue), and individual animal ID (red). In particular, a DJI Zenmuse X3 camera stream reshaped to $720 \times 720$ images is processed by a YoloV2-based Friesian cattle detector yielding RoI-annotated frames. These are transformed into an $5 \times 5$ occupancy grid map and also a set of spatio-temporal tracklets $\{T_0, T_1, ..., T_p\}$. The map is input to a tactical network based on AlexNet, which forms a dual stream navigation architecture together with a strategic network operating on a long-term explorational history memory $M$. Both streams are concatenated into a shallow navigation net that outputs a score vector $V$ based on which possible navigational actions $a \in \{N, W, S, E\}$ are selected. During operation selected actions are performed grounded in GPS and, in-turn, the positional history $M$ is updated. For individual ID each tracklet $T$ is rescaled and fed into an Inception V3 network up to the pool5 layer followed by a LSTM unit for temporal information integration mapping to a vector over the individual cattle IDs.

2. Hardware

We use the DJI Matrice 100 quadrotor UAV, which houses the ROS-enabled DJI N1 flight controller, as our base platform. It has been employed previously across various autonomous tasks [17, 8, 48, 24]. We extend the base M100 platform by adding an Nvidia Jetson TX2 mounted on a Connect Tech Inc. Orbitty carrier board to enable onboard deep inference via 256 CUDA cores under Nvidia’s Pascal\textsuperscript{TM} architecture. Also onboard is the DJI Manifold (essentially an Nvidia Jetson TK1) to decode the raw image feed from the onboard camera (the DJI Zenmuse X3 camera/gimbal) and to add further computational capacity. The X3 camera is mounted on rubber grommets and a 3-axis gimbal, which allows for rotor vibration isolation and independent movement of the flight platform for stabilised footage and programmatically controlled roll-pitch-yaw. A Quanum QM12V5AUBEC voltage regulation device was fitted to power non-conforming devices feeding off the primary aircraft battery. In addition, customised mounts were added for WiFi antennas to monitor the craft remotely (see Fig. 2).

3. Experimental Setup

Flights were performed over a two-week experimental period at the University of Bristol’s Wyndhurst Farm in Langford Village, UK on a consistent herd of 17 yearling heifer Holstein Friesian cattle satisfying all relevant animal welfare and flight regulations. Experiments consisted of two phases: (a) training data acquisition across 14 (semi)manual flights, and subsequent (b) conduction of 18 autonomous flights. Training data was acquired over two day-long recording sessions where manual and semi-autonomous flights at varying altitudes were carried out, recording video at a resolution of $3840 \times 2160$ at 30fps. The result was a raw dataset consisting of 37 minutes from 15 videos over 14 flights occupying 18GB. Overall 2285 frames were extracted from these video files at a rate of $1Hz$. First, after discarding frames without cattle, 3120 bounding boxes around individual cattle were labelled in the 553 frames containing cattle. Animals were also manually identified as ground truth for individual identification. Secondly, to produce ground truth for training the cattle detector, square sub-images were manually annotated to encompass individuals such that they are resolved at approximately $150 \times 150$ pixels. Fig. 3 illustrates the full pipeline.

Figure 1. Onboard System Overview. For individual Friesian cattle search and identification we propose an approach that utilises three deep convolutional neural architectures operating onboard a computationally-enhanced DJI Matrice 100 platform to achieve species detection (green), exploratory agency (blue), and individual animal ID (red). In particular, a DJI Zenmuse X3 camera stream reshaped to $720 \times 720$ images is processed by a YoloV2-based Friesian cattle detector yielding RoI-annotated frames. These are transformed into an $5 \times 5$ occupancy grid map and also a set of spatio-temporal tracklets $\{T_0, T_1, ..., T_p\}$. The map is input to a tactical network based on AlexNet, which forms a dual stream navigation architecture together with a strategic network operating on a long-term explorational history memory $M$. Both streams are concatenated into a shallow navigation net that outputs a score vector $V$ based on which possible navigational actions $a \in \{N, W, S, E\}$ are selected. During operation selected actions are performed grounded in GPS and, in-turn, the positional history $M$ is updated. For individual ID each tracklet $T$ is rescaled and fed into an Inception V3 network up to the pool5 layer followed by a LSTM unit for temporal information integration mapping to a vector over the individual cattle IDs.

2. Hardware

We use the DJI Matrice 100 quadrotor UAV, which houses the ROS-enabled DJI N1 flight controller, as our base platform. It has been employed previously across various autonomous tasks [17, 8, 48, 24]. We extend the base M100 platform by adding an Nvidia Jetson TX2 mounted on a Connect Tech Inc. Orbitty carrier board to enable onboard deep inference via 256 CUDA cores under Nvidia’s Pascal\textsuperscript{TM} architecture. Also onboard is the DJI Manifold (essentially an Nvidia Jetson TK1) to decode the raw image feed from the onboard camera (the DJI Zenmuse X3 camera/gimbal) and to add further computational capacity. The X3 camera is mounted on rubber grommets and a 3-axis gimbal, which allows for rotor vibration isolation and independent movement of the flight platform for stabilised footage and programmatically controlled roll-pitch-yaw. A Quanum QM12V5AUBEC voltage regulation device was fitted to power non-conforming devices feeding off the primary aircraft battery. In addition, customised mounts were added for WiFi antennas to monitor the craft remotely (see Fig. 2).

3. Experimental Setup

Flights were performed over a two-week experimental period at the University of Bristol’s Wyndhurst Farm in Langford Village, UK on a consistent herd of 17 yearling heifer Holstein Friesian cattle satisfying all relevant animal welfare and flight regulations. Experiments consisted of two phases: (a) training data acquisition across 14 (semi)manual flights, and subsequent (b) conduction of 18 autonomous flights. Training data was acquired over two day-long recording sessions where manual and semi-autonomous flights at varying altitudes were carried out, recording video at a resolution of $3840 \times 2160$ at 30fps. The result was a raw dataset consisting of 37 minutes from 15 videos over 14 flights occupying 18GB. Overall 2285 frames were extracted from these video files at a rate of $1Hz$. First, after discarding frames without cattle, 3120 bounding boxes around individual cattle were labelled in the 553 frames containing cattle. Animals were also manually identified as ground truth for individual identification. Secondly, to produce ground truth for training the cattle detector, square sub-images were manually annotated to encompass individuals such that they are resolved at approximately $150 \times 150$ pixels. Fig. 3 illustrates the full pipeline.
4. Implementation

Cattle detection and localisation is performed in realtime frame-by-frame using the YOLOv2 [38] CNN. The network was retrained from scratch on the annotated region dataset, consisting of 11,384 original and synthetic training images and associated ground truth labels. Synthetic instances were generated via image augmentation; stochastic combinations of rotation, cropping, affine transformations, etc. Model inference operates on $736 \times 736$ images obtained by cropping and scaling the source $960 \times 720$ pixel camera stream. As shown in Figure 1, this process yields a set of $m$ bounding boxes $B = \{b_0, b_1, ..., b_m\}$ per frame with associated object confidence scores. Inference on each of $n = 5$ sampled frames then produces a box-annotated spatio-temporal volume $\{B_0, B_1, ..., B_n\}$. Bounding boxes are associated across this volume by accumulating detections that are consistently present in cow-sized areas of an equally subdivided image. This method is effective for reliable short-term tracking due to distinct and non-overlapping targets, slow target movement and stable UAV hovering. The outputs are $p \leq n$ short individual animal tracklets reshaped into an image patch sequence $\{T_0, T_1, ..., T_p\}$, which forms the input to the individual identification network. In addition, the current frame $I$ is also abstracted to a $5 \times 5$ grid map $S$ encoding animal presence in the field of view of the camera (see Fig. 1).

Navigation activities aim at locating as many – themselves moving – individual animals as possible on the shortest routes in a gridded domain where a target counts as ‘located’ once the agent occupies the same grid location as the target. To solve this dynamic travelling salesman task (with city locations to be discovered on the fly), we use a dual-stream deep network architecture, as first suggested in our previous work [3]. The method computes grid based navigational decisions $\{N,W,S,E\}$ based on immediate sensory (tactical/exploitation) and historic navigational (strategic/exploration) information using two separate streams within a single deep inference network. As shown in the paper, this strategy can significantly outperform simple strategies such as a ‘lawnmower’ pattern and other baselines. To summarise the method’s operation, the sensory input $S$ is processed via a first stream utilising a basic AlexNet [22] design (see Fig. 1) because of its visual simplicity. A second stream operates on the exploratory history thus far, as stored in a long-term memory map $M$. This stores the agent’s present and past positions alongside animal encounters within the flight area of $20 \times 20$ grid locations. The agent’s starting position is fixed and $M$ is reset after $\delta$% of the map has been explored. Both these streams are concatenated into a shallow integration network that, as shown in Figure 1, maps to a SoftMax normalised likelihood vector $V$ of the possible navigational actions. During inference, the network selects the top-ranking navigational action from $\{N,W,S,E\}$ based on $V$, which is performed and, in-turn, the positional history $M$ is updated. For training, the entire two-stream navigation network is optimised via stochastic gradient decent (SGD) with momentum [36] and a fixed learning rate $e = 0.001$ based on triples $(S; M; V)$ using one-hot encoding of $V$ and cross-entropy loss. This unified model allows for the backpropagation of navigation decision errors across both streams and domains. For training, we simulate 10,000 episodes of 17 pseudo-randomly [28] placed targets in a $20 \times 20$ grid and calculate optimal navigation decisions $(S; M; V)$ by solving the associated travelling salesman problem. 10-fold cross validation on this setup yielded an accuracy of 72.45% in making an optimal next grid navigation decision and a target recovery rate of 0.26 $\pm$ 0.06 targets per move.

Individual identification based on an image patch sequence $\{T_0, T_1, ..., T_p\}$ is performed via an LRCN, first introduced by Donahue et al. [10]. In particular, as shown in Figure 1, we combine a GoogLeNet/Inception V3 CNN [44, 45] with a single Long Short-Term Memory (LSTM) [15] layer. This approach has demonstrated success in disambiguating fine-grained categories in our previous work [2]. Training of the GoogLeNet/Inception V3 network takes groups of $n = 5$ same class randomly selected RoIs, each of which were non-proportionally resized to $224 \times 224$ pixels. SGD with momentum [36], a batch size of 32 and a fixed learning rate $e = 0.001$ were used for optimisation. Once

![Figure 3. Training Data Annotation. Across all 15 videos gathered for training data acquisition, 553 frames contained cattle. Labelling bounding boxes around individual animals yielded 3120 cattle patches. Labelling 720 $\times$ 720 regions containing annotated cattle patches yielded the 1423 images annotated with cattle bounding boxes that form the base data for species detector training. Manual identification of the 17 cows, on the other hand, produced the base data for training the individual identification components.](image-url)
5. Results

We conducted 18 fully autonomous test flights at a low altitude (approximately 10m) above an area of 20 × 20 cells covering altogether 147 minutes. For all experiments, we ran the object detection and exploratory agency networks live. Note that the herd was naturally dispersed in these experiments and animals were free to roam across the entire field in and out of the covered area. Thus, only a few individuals were present and recoverable in this area at any one time. The median coverage of the grid was 70.13% with median flight time of approximately 8 minutes per experiment, and a median of 77 grid iterations per flight.

For offline analysis, the UAV saved to file one acquired 720 × 720 image at each exploratory agency iteration, yielding 1039 images. 99 of those images actually contained target cows that were hand labelled with ground truth bounding box annotations and identities according to the VOC guidelines [11]. A detection was deemed a successful true positive based on the IoU (\(\text{ov} \geq 0.5\)). Grounded in this, we measured the YOLOv2 detection accuracy to be 92.4%, where out of the 111 present animals, 2 were missed and 7 false positive nested detections occurred. We then tested separately, the performance of the single frame Inception V3 individual identification architecture (yielding 93.6% accuracy), where all ground truth bounding boxes (not only the detected instances) were presented to the ID component. In contrast, and as shown in Table 1, when identification is performed on detected RoIs only then the combined offline system accuracy is 91.9%.

<table>
<thead>
<tr>
<th># Sample Frames</th>
<th># Animal Instances</th>
<th>Detection on Sample Frames Accuracy (%)</th>
<th>Single Frame ID on Labelled RoIs Accuracy (%)</th>
<th>Combined Detection+ID Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1039</td>
<td>111</td>
<td>92.4</td>
<td>93.6</td>
<td>91.9</td>
</tr>
</tbody>
</table>

Table 1. Offline Performance Results.

For online autonomous operation, all computation was performed live in real-time onboard the UAV’s computers (DJI Manifold, Nvidia Jetson TX2). Figure 4 depicts various example flights with detailed annotations of flight paths, animal encounters and identification confidences. Across the 1039 grid locations visited during the set of experiments, the aircraft navigated centrally above a detected individual 18 times, triggering identification. These individuals represented a small subset (6 individuals) of the larger herd that tended to be more comfortable with the presence of the aircraft directly above them. Note that this mode of operation eliminates the problem of clipped visibility at image borders, minimises image distortions, optimises the viewpoint, and exposes the coat in an orthogonal view. For triggered identification, we store LRCN confidence outputs after processing up to 5 same class patches \(T\) to compare performance differences between single view and multi-view identification. The respective overall results are given in Table 2. Notably, across the small sample set, the model performed perfectly.

Table 2. Online Identification Results.

<table>
<thead>
<tr>
<th># Samples</th>
<th>LRCN Identification Accuracy (%)</th>
<th>Single Frame Identification Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>18</td>
<td>100</td>
<td>94.4</td>
</tr>
</tbody>
</table>

6. Conclusion

The experiments conducted confirm demonstrable identification robustness of the approach. This gives rise to future agricultural automation possibilities with potential implications for animal welfare and farm productivity. Beyond farming, the concept of autonomous biometric animal identification from the air as presented opens up a realm of future applications in fields such as ecology, where animal identification of uniquely patterned species in the wild (e.g., zebras, giraffes) is critical to assessing the status of populations. Future work will investigate the effect of varying the UAV altitude and consequently, varied environment visibility, target resolution, etc.
References


