Object Detection for Cattle Gait Tracking

John Gardenier¹, James Underwood¹, Cameron Clark²

Abstract—Lameness in cattle is a health issue where gait is modified to minimise pain. Cattle are currently visually assessed for locomotion score, which provides the degree of lameness for individual animals. This subjective method is costly in terms of labour, and its level of accuracy and ability to detect small changes in locomotion that is critical for early detection of lameness and associated intervention. Current automatic lameness detection systems found in literature have not yet met the ultimate goal of widespread commercial adoption. We present a sensor configuration to record cattle kinematics towards automatic lameness detection. This configuration features four Time of Flight sensors to view cattle from above and from one side as they exit an automatic rotary milking dairy. Two dimensional near infrared images sampled from 223 cows passing through the system were used to train a Faster R-CNN to detect hooves (F1-score = 0.90) and carpal/tarsal joints (F1-score = 0.85). The depth images were used to project these detected key points into Cartesian space where they were tracked to obtain individual trajectories per limb. The results show that kinematic gait features can be successfully obtained as a first and important step towards objective, accurate, automatic lameness detection.

I. INTRODUCTION

Lameness in cattle is observed as impaired motion, or non-standard gait or posture due to painful disorders in the locomotor system [1], [2]. Lameness in dairy cattle is a prevalent health issue impacting animal welfare and economic performance and is most often caused by hoof lesions [3]. Detecting lameness systematically is a crucial step in reducing lameness on dairy farms. Since systematic manual methods of detecting lameness are time consuming and therefore often neglected [4], and increasing herd sizes [5], [6] are resulting in less farmer time available per cow, it is proposed that automating the detection task will improve lameness management, including prevention and treatment. In turn, animal welfare will be improved and monetary losses from lameness decreased due to early detection and intervention.

Multiple methods exist, both manual and automatic, for detecting lameness. In a manual locomotion scoring system (e.g. [7], [8]) the locomotion of cattle is scored on an ordinal scale by human scorers who watch for specific locomotion traits [9], resulting in a lameness score, where a low number indicates mild lameness and a high number indicates severe lameness. Performing locomotion scoring requires time and skilled scorers, the lack of which results in poor lameness management on dairy farms [4]. In contrast to manual lameness detection, automatic lameness detection could provide an objective, consistent lameness assessment at higher temporal resolution, while potentially being able to detect smaller changes in gait. While automatic lameness detection methods exist in literature using a range of sensing modalities and lameness modelling techniques, none have yet led to widespread commercial adoption [1], in part due to robustness and performance issues. Coupled with recent advances in low cost, high quality Time of Flight (ToF) sensors, and state-of-the-art deep learning methods, we present a sensor configuration to record cattle kinematics towards objective, accurate, automatic lameness detection.

This manuscript provides:

- sensor and frame design suitable for capturing 2D and 3D imagery from top and one side of cattle
- results of hoof and carpal/tarsal joint detection and tracking
- diagrams showing gait of individual cattle, and derived gait parameters including step overlap and abduction

II. RELATED WORK

Automatic lameness detection has been performed by measuring cattle kinetics, kinematics, and by indirect measures [9]. In the kinetic approach, the forces exerted by the hoof on the ground are measured by force plates, such as in [10] where individual hoof strikes are extracted from two longitudinal force plates cattle walk over. Measuring cattle kinematics, that is the motion of specific body parts temporally and spatially, has shown better results than purely force based approaches. Lameness has been predicted from back curvature monitored with side-on 2D footage [11] and with top-down 3D cameras [12], [11], [13]. Lameness has also been predicted from hoof placement [14], [15], ankle location [16], and cattle walking speed [17] extracted from 2D side-on footage. The Gaitwise system uses a pressure mat to obtain the position and force of each hoof strike [18], [19], where the most important metrics are related to position of the hoof strike, not the force, meaning that a kinematic approach is preferred over a kinetic approach for lameness detection. Accelerometers attached to the four limbs have also been used in kinematic studies [20]. Indirect approaches include on-cow sensors such as accelerometers and GPS loggers, which can be used to measure behaviour (e.g. lying, standing, walking time) [21], and off-cow sensors such as milking and feeding parameters [22].

Kinematic approaches perform better than kinetic approaches as data is collected during the entire gait, including when a limb is off the ground, giving a better characterisation
of gait than purely the forces and times of hoof placement. Indirect approaches suffer from other health issues corrupting the link between lameness and the recorded data to a greater extent than for the more direct kinetic or kinematic approaches. In order to be able to predict multi-class lameness at state-of-the-art performance levels, we chose to design a sensor configuration capable of recording cattle kinematics, both of the limbs and of the back region.

To capture gait kinematics, key point detection is used to extract specific locations of interest from imagery. In automatic lameness detection literature, these are body parts of the locomotor system. In previous work the body parts of interest were hook bones and spine [12], and spine curvature and head angle [23]. Both studies use hand-engineered features from depth and colour imagery to extract the desired body parts. In other computer vision applications, learned features have shown to perform better than hand engineered features. Key points have been extracted for human gait recognition using Convolutional Neural Networks (CNNs) represented as confidence maps [24], [25] and as $x$, $y$ pixel coordinates [26]. The features extracted by CNNs can be used for a variety of tasks, e.g. classification, detection, and regression. A recent object detection framework, Faster R-CNN [27], has been used in this manuscript to detect key points. While other detection frameworks have been tuned for better key point detection performance, the Faster R-CNN framework is very common and easy to implement.

Individual key point detections are linked together to form tracks in space and time using multi-object tracking algorithms. Two key components of tracking are data association and position estimation filters, which can be performed online or offline. Examples of data association are joint probabilistic data association [28] and multiple hypothesis tracking [29]. Association between objects in subsequent frames has also been learned from CNN features in [30]. In this manuscript, a standard Kalman filter has been used for position estimation, coupled with a correspondence matrix with 3D Euclidean distance for data association [31]. A recent literature review of multi-object tracking can be found in [32].

III. EXPERIMENTAL SETUP

This section describes our initial steps towards an automatic, kinematic cattle lameness detection system, including sensor choice and configuration, how data were collected, and the pipeline from sensor data to key point detection to key point tracking.

A. Sensor choice

1) Sensor choice: Four types of 3D sensors were considered to capture cattle kinematics: LiDAR; Time of Flight (ToF); Structured Light (SL); and Stereo (infrared or colour). Representative sensors for each type were assessed for cost, availability, and suitability for use in an automatic lameness detection system. One of the restrictions seen in previous research with mainly SL sensors was that the sensors could only operate in the dark, as sunlight would overpower the light emitted by the active sensor. Stereo sensors can suffer from the opposite situation where there is not enough environmental light for the passive cameras. Informal tests in a representative environment were conducted with two SL sensors, the Microsoft Kinect-v1 and Intel Realsense SR300, and two stereo sensors, the Point Grey Bumblebee XB3 and Intel Realsense R200. The SL sensors did not produce any data in bright sunlight and the stereo sensors were found to have poor dense 3D reconstruction quality.

Velodyne HDL-64 and VLP-16 LiDAR sensors, and Microsoft Kinect-v2 ToF sensors were tested at the dairy. The 64-line Velodyne was placed 8 m to the left side of passing cattle and scanned vertically, as seen in the vertical stripes on the cow in Fig 1b. The 16-line Velodyne was placed 1.3 m to the left of passing cattle and scanned horizontally for more detail of the limbs, as seen in the horizontal stripes on the limbs and floor in Fig. 1b. The Kinect-v2 ToF sensor was tested in a variety of lighting conditions, at multiple distances and orientations with respect to the cattle. The quality of 3D and intensity data received from the Kinect-v2 was satisfactory, with a higher resolution than the LiDAR sensors could provide. While the ToF sensor produced more stray, erroneous points than the LiDAR sensors, the number of points in the scene is an order of magnitude greater for ToF than for LiDAR as can be seen in Fig. 1. A theoretical advantage of LiDAR is that it is completely impervious to the illumination conditions and operates well in full sunlight and complete darkness, whereas the Kinect-v2 may fail in extremely bright environments. This was not, however, observed in any of our tests. Interference between multiple Kinect-v2 sensors was tested informally and found only to be an issue when the sensors are facing one another, or have a large overlapping scene.

2) Dairy and sensor configuration: The prototype used in this manuscript had four Kinect-v2 sensors mounted to an inverted L-shaped frame as seen in Fig. 2. This configuration allowed for two viewpoints as cattle walked past upon exiting the rotary milker. Two sensors were placed at the end of the arm looking down on the cow’s back and head, while the other two sensors were placed on the vertical post looking across at the cow’s limbs. The limbs were partially obscured by the near fence. Having two viewpoints allowed data to be collected from both the head/back region and the limbs, which are most often used in determining lameness. At both viewpoints, one sensor was tilted towards the cattle entering the region of interest, while the second sensor was tilted towards the cattle exiting the region of interest. This sensor configuration did not capture the far side of the cow; however, such a configuration with a full archway frame as in [33] would be more intrusive and complex. The lower, most important part of all four limbs was visible from one side of the cow.

The sensor system location in the dairy was ideal for automatic lameness detection for a number of logistical and cow presentation reasons. It is common to monitor cows for lameness after milking so that udder fill does not influence gait. The cattle walked past one by one as they stepped off
the automatic rotary milking dairy, and they were motivated to walk at their own pace towards a sorting gate giving them access to automatic feed hoppers. Two fences guided cattle down this 3.6 m path, ensuring that the cattle walked in a straight line. The path length allowed a complete gait cycle to be observed. To the left of the cattle path, there was ample room place sensors to span the full 3.6 m path with their field of view. Although the Kinect-v2 sensors performed well in full sunlight tests, this area was also advantageous because it was covered by a high open roof, mitigating potential direct sunlight overpowering the active Time of Flight sensors.

A total of 223 cows were recorded in May 2017 at The University of Sydney’s Corstorphine Farm in Camden, Australia. Infrared illumination and depth imagery were logged from the Kinect-v2 cameras at 30 Hz at the native resolution of 512 x 424 pixels for both channels. For manual lameness scoring as ground truth, a 1280 x 1024 colour camera recording at 30 Hz was placed 8 m from the left side of the cattle. Lameness scoring was performed from footage by a single observer using the Dairy Australia 4-level lameness scale [8]. These scores will be used as ground truth in future work, however were not required in this manuscript.

B. Key point detection

Specific body parts, or key points, were detected using a Region based Convolutional Neural Network framework. The Faster R-CNN framework [27] was used with the VGG16 feature extractor as implemented in [34]. During training, the input to the R-CNN framework were images (3 channels for the VGG16 feature extractor) and corresponding bounding boxes with class label. Once a network is trained, an input image results in detections as output, quantified by a bounding box, class, and class probability score. In this manuscript, separate single class networks were trained to form a binary detector per class, and the input to the feature extractor was a three channel image made by simply repeating the single channel Kinect-v2 near infrared frame.

1) Labelling: Labelling was performed with [35] on 74 near infrared images randomly sampled from imagery of
the two lower sensors when cattle are passing through the system. Unobscured hoof tips, hoof heels, and carpal and tarsal joints were labelled with a single click. The corresponding label class was distinct for each limb (e.g. front/rear, left/right). A virtual hoof location in the image was constructed halfway between a hoof tip and hoof heel, with a vertical offset of 2 pixels to ensure the location was on the hoof. The labels of all four limbs were merged to form a single carpal/tarsal class and single hoof class for training and detection, of which the quantities can be found in Table I. Square bounding boxes were centred on the point labels, resulting in bounding box labels. Two sizes of bounding boxes, 20 and 50 pixels, were evaluated separately.

2) Training: The 74 images containing both background images and images with cattle walking past were split randomly into 5 folds, each containing 14 or 15 images. Each fold was used once as a test set, with the remaining 4 being split randomly into a training set of 45 images and a validation set of 14 or 15 images. The weights of a VGG16 feature extractor trained on the ImageNet dataset were used as a starting point for training [36], and were then fine tuned to the task of detecting hooves and carpal/tarsal joints. The feature extractor weights were fine tuned for 5000 iterations for each of the 5 training sets for each class and bounding box label size. During training data augmentation was performed by randomly flipping and scaling the image per epoch, and a true positive detection was assumed when the bounding box label and detection had an intersection of union greater than 0.2.

3) Evaluation: The key point detection performance was measured by the F1-score, which is the harmonic mean of precision and recall commonly used to quantify binary classification performance. The F1-score is bounded between 0 (poor performance) and 1 (high performance), where higher scores are achieved when both precision and recall are high. For the results presented in this section, two sets of independent networks were trained for hoof detection and for tarsal/carpal joint detection. While the body parts of the four limbs were annotated individually, i.e. left front carpal joint, right front carpal joint, the networks were trained to detect all carpal/tarsal joints irrespective of limb. The visual similarity between the hooves and carpal/tarsal joints of the four limbs might be too great for visual classification into separate classes, and that complexity is unnecessary.

The detection threshold was determined on the validation data as the average across the data splits of the optimum detection threshold for highest F1-score at 5000 iterations. The reported F1-score was evaluated by using the detection threshold on the 5 test sets and averaging the resulting F1-score.

Localisation accuracy was evaluated by using the Mean Absolute Error (MAE) on the test set between the centres of detections and corresponding bounding box labels when they are within half of the bounding box label size of each other. The MAE was calculated both for 2D error and 3D error.

Once the detection performance was analysed on the 5 test sets, all labelled images were used to train a final network which was trained for 20000 iterations.

C. Tracking

Tracking was performed for the hooves detected from a single lower sensor during a single cow passing with the final hoof detector network. The detections were tracked over the distance between the two vertical posts of the near fence in Fig. 3 to avoid the tracking filters having to cope with missing detections when obscured behind the vertical post.

1) Projection to 3D space: By performing a standard pinhole camera model intrinsic calibration for each Kinect v2 sensor, each depth value in the depth map can be projected into 3D space. To obtain a depth value for each detection in 2D, the most frequently occurring depth value (the mode) within a 21 pixel square centred around the detection centroid was used, along with the coordinates of the detection centroid. The value of 21 pixels was obtained via trial and error: if the value was too small, detection centres that missed the hoof were assigned a background depth, while too large a value resulted in dominance of the background depth regardless of the centroid accuracy. Changing this square size controls errors in depth, but not the 2D centroid of the detection. All further analysis is performed in 3D space.

2) Data association and filtering: To be able to extract gait parameters from the 3D key point detections, the detections in the current frame need to be associated to corresponding detections in previous and next frames. This is of particular interest to key points on limbs such as hooves that are detected as part of the same class in this manuscript. For example, a nose key point is a unique class and therefore does not require data association to distinguish multiple instances but would however handle missing and/or noisy detections. Data association was performed using the 3D Euclidean distance between previous tracks and new detections, and the correspondence matrix [31]. Additionally, the newly associated detections were used as an observation in a Kalman filter per track.

3) Track to limb association: Individual tracks were associated to one of the four limbs by matching hoof placement timestamps and geometry to a cow gait cycle. The first and last hoof placement geometry provided an additional source of information that was used to further improve limb disambiguation.

4) Evaluation: Tracking performance was evaluated with the MAE of the Euclidean distance in 3D space between labels and detections. Undetected hooves do not negatively impact MAE. The 421 labels for tracking evaluation in 235 frames were not used to train the key point detectors.

<table>
<thead>
<tr>
<th>TABLE I: Labelled key points</th>
</tr>
</thead>
<tbody>
<tr>
<td>Key point</td>
</tr>
<tr>
<td>Carpal/tarsal joint</td>
</tr>
<tr>
<td>Hoof tip</td>
</tr>
<tr>
<td>Hoof heal</td>
</tr>
<tr>
<td>Hoof</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Hoof 78</th>
<th>Hoof heal 93</th>
<th>Hoof tip 90</th>
<th>Carpal/tarsal joint 77</th>
</tr>
</thead>
</table>
Fig. 3: Detection of carpal/tarsal joints in orange and hooves in blue. Crosses are label centroids, bounding boxes are detections. Left rear hoof was neither labelled nor detected due to hoof tip being obscured. Best viewed in colour.

TABLE II: Detection results showing binary classification performance

<table>
<thead>
<tr>
<th>Key point</th>
<th>Label size [px]</th>
<th>Detection threshold</th>
<th>F1-score</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hoof</td>
<td>20</td>
<td>0.33</td>
<td>0.86</td>
<td>0.90</td>
<td>0.83</td>
</tr>
<tr>
<td>Hoof</td>
<td>50</td>
<td>0.22</td>
<td>0.90</td>
<td>0.92</td>
<td>0.89</td>
</tr>
<tr>
<td>Carpal/tarsal</td>
<td>20</td>
<td>0.22</td>
<td>0.72</td>
<td>0.75</td>
<td>0.69</td>
</tr>
<tr>
<td>Carpal/tarsal</td>
<td>50</td>
<td>0.14</td>
<td>0.85</td>
<td>0.83</td>
<td>0.87</td>
</tr>
</tbody>
</table>

*aAveraged over 5 test sets

IV. RESULTS

This section presents the detection and localisation performance for carpal/tarsal joints and hooves, along with tracking performance for hooves for a single non-lame cow passing through the system. Tracked hoof positions and hoof placement plots are shown for both a non-lame cow and a lame cow.

A. Key point detection

A single frame with labels and detections for hooves and carpal/tarsal joints is shown in Fig. 3. Table II shows the detection results as an average over the five test sets. The localisation accuracy is given in Table III in 2D and 3D as average Mean Absolute Error (MAE) over the same five test sets.

B. Key point tracking

Fig. 4a shows a side view of the tracked hoof positions for a non-lame cow. The tracking accuracy between detections and labels for hooves was evaluated as the 3D Mean Absolute Error (MAE) over the same five test sets.

TABLE III: Localisation results showing Mean Absolute Error (MAE) of detection centroids

<table>
<thead>
<tr>
<th>Key point</th>
<th>Label size [px]</th>
<th>MAE 2D [px]</th>
<th>MAE 3D [cm]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hoof</td>
<td>20</td>
<td>4.24</td>
<td>2.91</td>
</tr>
<tr>
<td>Hoof</td>
<td>50</td>
<td>4.82</td>
<td>3.39</td>
</tr>
<tr>
<td>Carpal/tarsal</td>
<td>20</td>
<td>5.02</td>
<td>4.51</td>
</tr>
<tr>
<td>Carpal/tarsal</td>
<td>50</td>
<td>8.15</td>
<td>8.18</td>
</tr>
</tbody>
</table>

*aAveraged over 5 test sets

TABLE IV: Tracking results showing Mean Absolute Error (MAE) of detection centroids after tracking

<table>
<thead>
<tr>
<th>Key point</th>
<th>Label size [px]</th>
<th>MAE 3D [cm]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hoof</td>
<td>20</td>
<td>3.94</td>
</tr>
<tr>
<td>Hoof</td>
<td>50</td>
<td>7.31</td>
</tr>
</tbody>
</table>

longitudinal velocity is lower than 0.4 ms\(^{-1}\), as can be found in Fig. 4b. Both tracking and hoof placement results are presented for a severely lame cow in Fig. 5.

V. DISCUSSION

The system was able to accurately detect hooves (F1-score=0.90) and carpal/tarsal joints (F1-score=0.85). F1-scores for the hoof class were higher than for the carpal/tarsal class. This could be explained by the carpal and tarsal joints being merged into a single class while their appearance is more variable than for the four hooves. Separating the tarsal and carpal joints into two classes (front/rear) may improve the F1-score as they are qualitatively distinct. The VGG16 feature extractor is able to represent much more variation than is presented in the 74 training images, which may result in over fitting. In future work performance will be improved by increasing the size of the training data set by labelling more images and by using images from different farm sites. Additionally, data association and tracking would be simplified if a four (left/right, front/rear), or even two (front/rear) independent networks were trained for the four hooves and carpal/tarsal joints.

A larger bounding box label size (20 px versus 50 px) resulted in a higher F1-score for both hoof (F1-scores=(0.86, 0.90)) and carpal/tarsal (F1-scores=(0.72, 0.85)) classes. However, as bounding box label size increased, a larger 2D localisation error is observed for hooves (MAE from 4.2 px to 4.8 px) and carpal/tarsal joints (MAE from 5.0 px to 8.2 px). The corresponding 3D localisation error also increased for hooves (MAE from 2.9 cm to 3.4 cm) and carpal/tarsal joints (MAE from 4.5 cm to 8.2 cm). The effect of a larger bounding box label size compared to a smaller bounding box label size with equal centroids was an increase in the intersection of union and hence F1-score, while the exact position of the centroid of these larger detections became less accurate.

Tracking performance was evaluated by MAE in 3D for the hooves of the cow passing depicted in Fig. 4. For both 20 pixel (MAE = 3.9 cm) and 50 pixel detections (MAE=7.3 cm) we saw an increase compared to the 3D MAE detection results (MAE=(2.9 cm, 3.4 cm)). This can be explained by
(a) Side view of hoof trajectories where hooves enter scene from the right. Left hooves show a complete stride and are placed on ground twice while right hooves enter scene mid-stride and have a single hoof placement.

(b) Top-down view of hoof placement on ground, extracted from hoof trajectory where longitudinal velocity is less than 0.4 ms$^{-1}$. Proximity of rear vs. front hoof placement is shown for left hooves.

Fig. 4: Tracking results for a non-lame cow. Cow motion is from right to left. For both left and right limbs, rear hoof is placed close to front hoof.

(a) Side view of hoof trajectories where hooves enter scene from the right. Right hooves show a complete stride and are placed on ground twice while left hooves enter scene mid-stride and have a single hoof placement.

(b) Top-down view of hoof placement on ground, extracted from hoof trajectory where longitudinal velocity is less than 0.4 ms$^{-1}$. Proximity of rear vs. front hoof placement is shown for left hooves.

Fig. 5: Tracking results for a severely lame cow. Cow motion is from right to left. For both left and right limbs, rear hoof is placed further away from front hoof than for non-lame cow in Fig. 4.
the influence of the Kalman filter in the tracking, which smooths incoming detections.

The tracked hoof positions can be used to calculate a variety of gait parameters. For a non-lame cow one expects the rear hoof to be placed at the same location as the front hoof for each side. Abduction is a measure of the lateral distance between front hoof and rear hoof placement, while step overlap is the equivalent measure in longitudinal direction. Abduction and step overlap have been annotated for the left limbs in Figs. 4b and 5b for a non-lame and lame cow respectively. Both abduction and step overlap are higher for the severely lame cow (abduction = 26 cm, step overlap = 10 cm) compared to the non-lame cow (abduction = 10 cm, 5 cm, step overlap = 3 cm, 4 cm). The presence of unfamiliar sensors to the left of the cow likely influenced the abduction of the left pair of limbs in Fig. 5b. In future studies equipment will be installed for a period before recording in order for the cattle to grow familiar to their appearance. Both abduction and step overlap have been used together with other gait parameters to predict lameness [18]. Having successfully extracted hoof placement positions, many more gait parameters will be able to be computed in the future in order to predict lameness. Additionally, more body parts than hooves and carpal/tarsal joint will be tracked in order to represent gait dynamics to a greater extent.

VI. CONCLUSION

This manuscript presented a perception system that is suitable for automatic lameness detection. Kinect-v2 3D sensors were placed above and alongside cattle exiting an automated milking system, and recorded gait over a length of 3.6 m. A Faster R-CNN with VGG16 feature extractor was trained on 74 labelled near infrared images and provided accurate detections of hooves and carpal/tarsal joints. Hoof detections were projected to 3D and tracked to form 4D trajectories (space and time) of each of the four hooves. Gait parameters were calculated from the trajectories, including abduction and step overlap. Future work includes expanding the number of classes that can be detected, and using the obtained tracks to predict lameness score.

ACKNOWLEDGMENT

This work is funded by Dairy Australia Ltd with support from the Australian Centre for Field Robotics and the Dairy Science Group at The University of Sydney.

REFERENCES


