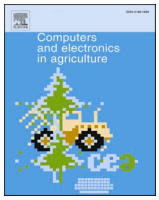
Computers and Electronics in Agriculture 229 (2025) 109812

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

journal homepage: www.elsevier.com/locate/compag



Original papers

Tag 'n' Track: Tackling the validation challenge in animal behaviour studies through automated referencing with ArUco markers

Serge Alindekon [a], Jana Deutsch [a,c], T. Bas Rodenburg [b], Jan Langbein [c], Birger Puppe [c,d], Helen Louton [a,e,*]

[a] Animal Health and Animal Welfare, Faculty of Agricultural and Environmental Sciences, University of Rostock, Justus-von-Liebig-Weg 6b, 18059 Rostock, Germany
[b] Animals in Science and Society, Faculty of Veterinary Medicine, Utrecht University, Yalelaan 2, 3584 CM, Utrecht, the Netherlands
[c] Behaviour and Welfare, Research Institute for Farm Animal Biology (FBN), Wilhelm-Stahl-Allee 2, 18196 Dummerstorf, Germany
[d] Behavioural Sciences, Faculty of Agricultural and Environmental Sciences, University of Rostock, Justus-von-Liebig-Weg 7, 18059 Rostock, Germany
[e] Chair of Animal Welfare, Ethology, Animal Hygiene and Animal Husbandry, Department of Veterinary Sciences, Faculty of Veterinary Medicine, LMU Munich, Veterinaerstr. 13/R, 80539 Munich, Germany

ARTICLE INFO

Keywords:
Automated Behavior Analysis
Animal Tracking Technology
Validation
Pattern Recognition
Precision Livestock Farming

ABSTRACT

Technological advances promise to greatly assist the study of animal behaviour, but the validation of these technologies is often neglected due to its tedious and labour-intensive nature. This paper addresses the challenges of manual annotation in validating technological tools for animal behaviour research. We detail the implementation and effectiveness of a computer vision method that automatically annotates animals within various regions of interest (ROIs). This method uses ArUco markers, open-source visual markers with a grid pattern, fitted onto vests worn by the animals. To validate this method, we used 245 10-minute videos capturing animals' visits to key resources, using a mobile barn housing twenty-one chickens. Our method generates annotated videos, revealing unique IDs of individuals and timestamps marking their presence in ROIs. Compared with traditional human observation, our method performed excellently: Spearman's correlation ($\rho = 0.96$, $p < 0.01$), 92.83 % sensitivity, 99.93 % specificity, 99.08 % accuracy, 98.77 % precision, and a 95.28 % F1-score. All recordings were annotated automatically in 40.96 h, saving 82.72 % of the time compared to the 222.84 h required for manual annotation. The proposed ArUco marker-based tracking method is easy to set up, based on open-source technology, and accessible to researchers without advanced programming skills. This method has the potential to replace or complement manual annotation, simplifying the validation of new technologies for automated individual tracking.

1. Introduction

How much can we rely on the emerging technologies that offer innovative approaches to monitoring animal behaviour? Given the proliferation of technological tools recently introduced to study animal behaviour, it becomes crucial to question their reliability and validity. The need for validation of such tools prior to an observational or experimental study, as Siegford et al. (2023) emphasised, should not be underestimated.

The study of animal behaviour, defined as the systematic observation of the coordinated responses of animals to various stimuli, whether internal or external (Levitis et al., 2009), is essential. Not only does it enable us to better understand animals' sensitivity and reactivity to their physical and social environment, but it also aids the identification of essential factors that influence their preferences, welfare and productivity. In the study conducted by AHAW (2015) on perch design, a comprehensive analysis of individuals' behaviour revealed that various factors, including the material, shape, length, and width of the perch, exert a significant influence on hens' preferences and choices, emphasising the crucial role animal behaviour studies play in understanding their needs and optimising environmental conditions. This is an example of applied animal behaviour research informing policy and supporting the livestock industry. In fundamental research, careful observation of animal behaviour offers valuable insights into aspects such as social interactions (Collias and Collias, 1996; Gómez et al., 2022).

As more and more technologies are used in such essential studies, a

* Corresponding author.
E-mail address: h.louton@lmu.de (H. Louton).

https://doi.org/10.1016/j.compag.2024.109812
Received 15 January 2024; Received in revised form 4 November 2024; Accepted 8 December 2024
Available online 31 December 2024

problem emerges: some tools may provide incorrect information because their validity has not been adequately verified. Many researchers neglect this crucial step before adopting new tracking tools. For example, almost half of the studies using RFID to track poultry have not tested the tool's validity (Alindekon et al., 2023). This highlights significant concerns about the reliability of the data collected using these tracking technologies. It is difficult to ensure an accurate interpretation of animal behaviours without adequate validation — an assessment of whether a monitoring tool actually measures the behaviours it is supposed to measure in a reproducible and reliable manner.

The lack of thorough validation of animal behaviour monitoring tools can be attributed to the inherent labour-intensive nature of manual annotation by humans, either in real-time or on-screen, to evaluate the effectiveness of these tools. Manual annotation can be exhausting, requiring considerable labour, resources and time. Sales et al. (2015) manually analysed 256 h of video to validate an RFID technology, describing the experience as "time-intensive".

Although manual annotation undeniably remains the benchmark for validation, its lack of practicality, sometimes rudimentary nature, and challenges have led us to propose an alternative method. This method is intended to substitute or complement manual annotation when constraints make it difficult or impossible. Our overall aim in this paper is to present an automated tracking technology based on visual markers that is open-source, reliable, and adaptable to many animal species with appropriate adjustments, requires no prior programming skills, and is intended to serve as a tool for validating other tracking technologies.

Among the tracking methods based on pattern recognition, ArUco markers are a popular option. They are widely used in various fields, such as video editing, augmented reality in cinematography, camera calibration (Čepon et al., 2023), robot navigation and localisation, autonomous vehicles (Blachut et al., 2022) and drones (Siki and Takács, 2021). ArUco markers are a simplified version of QR codes, characterised by their streamlined appearance (Fig. 1).

These markers have been specially designed to be easily detected and tracked, among other visual elements (Garrido-Jurado et al., 2014, 2016). As soon as a marker is identified during image analysis, the algorithm provides the identifier, which can be associated with the identification timestamp or location (e.g., Eagan et al., 2022), enabling precise tracking and identification of the marked object (Wubben et al., 2019).

Fig. 1. This shows what an ArUco marker looks like. This is a marker from a set to which ID = 1 has been pre-assigned. These patterns are made freely available from one of the many online ArUco marker generators to create markers with fixed IDs.

ArUco markers could encourage the validation of other animal-tracking technologies for a number of reasons. Firstly, their simplicity and accessibility are major advantages. Adopting ArUco markers does not require advanced programming skills; they are based on existing computer vision techniques (Garrido-Jurado et al., 2014, 2016). In addition, these markers are characterised by their accuracy, robustness and versatility. They can detect fine movements ranging from 0.1 to 0.01 mm (Siki and Takács, 2021). They remain effective even when the marker is tilted by up to 70 degrees (Ferrão et al., 2018; Koeda et al., 2018), or for fast-moving objects of up to 70 km/h (Blachut et al., 2022). Their size is not a constraint as long as they remain visible in the videos. They have been used successfully to track insects as varied as ants (Sclocco et al., 2021) and cockroaches (Othayoth et al., 2022).

Our paper has two specific objectives:

(1) To detail the implementation of a 3D ArUco marker system for reliable automated annotation of animal presence within a region of interest; this can then be used to validate other animal tracking technologies that are typically used for tracking animal presence, as long as the 3D-marker does not interfere with their functionality. We assigned a unique ID to each animal by exploiting a pre-existing pattern identification system. Using chickens housed in a mobile barn as a model, we demonstrate the effectiveness of this method, particularly for analysing the use of key resources required to meet physiological and behavioural needs and to provide opportunities for improving welfare in poultry.
(2) To validate the proposed 3D ArUco marker by comparing it with human annotation, widely recognised as a standard reference. Quantitative assessments were conducted to evaluate the performance of our proposed technique compared to human reference. These assessments included calculating key metrics such as sensitivity, specificity, accuracy, and precision. This second objective aims to establish a trustworthy reference for future studies and ensure the validity of our method. It is also meant to serve as a model or case study for other research, including validation efforts, reinforcing that the primary intention of this paper is to provide a reliable validation tool rather than a comprehensive behavioural study.

2. Materials and methods

2.1. Ethical considerations

The University of Rostock's internal animal welfare committee reviewed and approved the research protocol. The study commenced upon receiving the necessary approvals (No. AZ 7221.3-18196_23) from the ethics committee and the veterinary office in Mecklenburg–Pomerania, Germany.

2.2. Experimental setup

2.2.1. Animals, housing and management

Twenty-one chickens of mixed lines and sexes were included in this study. Four of them were of the Lohmann breed (1 white Lohmann Selected Leghorn and 3 of Lohmann Brown) and were 45 weeks old; the remaining consisted of 16 females and one male of the Vorwerk breed, a traditional dual-purpose breed of chicken, all of which were 68 weeks old. This diversity is essential to reflect lineage and sex variability in resource visits.

The study was carried out at the "Friedrich Harms" Animal Experiment Station of Rostock University in Dummerstorf, Germany, using a customised version of the mobile poultry barn ROWA 200 v4.0 (M&Z

Manufaktur GmbH, Germany). The mobile barn has two areas: a main compartment and a winter garden.

The main compartment (6.04 m × 2.80 m × 2.00 m) was equipped with the wooden Europa Nest®, with six nests in two rows. These nests measured 28 cm × 38 cm × 25 cm × 50 cm (length × width × height1 × height2). The animals had access to two types of perches. One was a cylindrical metal rod, 2.25 m long, placed above the drinking line. The other was a two-tiered L-shaped perch made of pine wood and inclined at 38 degrees. Each level of this perch was 1.95 m long. The upper level was positioned 60 cm above the floor, while the lower level was 18 cm from it. The feeder was linear, made of plastic, measuring 20 cm × 200 cm, and accessible from both sides for the chickens. The drinker line comprised three bowls and 12 nipples. LED lighting has also been installed, programmed to provide 16 h of light daily, starting at 5:30 am.

The winter garden (6.04 m × 2.48 m × 2.00 m) was a semi-open compartment shielded by a partially transparent tarpaulin with netting. It was an indoor walking area with an outdoor climate that allowed the animals to perform natural behaviours like scratching while benefiting from natural light and fresh air and still being protected. The floor was covered with wood shavings. The winter garden was accessible through a 120 cm × 35 cm pophole with automatic opening and closing. The winter garden also featured an alfalfa bale, a tray with mineral blocks, scattered grit and oyster shells as supplements, and a box filled with a mix of zeolite rock powder and wood shavings for the chickens' dust baths. We occasionally scattered grains in the winter garden throughout the day to promote foraging behaviour.

The hens had access to 3 kg of feed (PANTO® LMK Legemehlkorn, Germany), distributed each morning manually. Water was available *ad libitum*. The animals had been vaccinated against Newcastle disease, and their well-being was visually checked every day. Poultry lice traps were set, and hen droppings were sampled regularly for parasite control. Production, health and environmental data were regularly monitored and amended as required. The winter garden was accessible to the chickens from 9:00 am to 8:15 pm. The animals were gradually acclimatised for 15 days to the main compartment, the winter garden and body-worn equipment.

Throughout this paper, we refer to seven resources collectively as "key resources". These are: drinker, enrichments within the winter garden, feeder, indoor litter area, pophole allowing access to the winter garden, and both metal and wooden perches. These are considered as such because they are important for the chickens to meet their physiological and behavioural needs; they also offer them behavioural opportunities, enhancing their welfare. During data collection, the animals preferred laying eggs in the indoor litter area rather than in the provided nests. Consequently, we considered that area as a "key resource" over the actual nest. The listed key resources served as an effective model for our study because they are generally the focus of research where technologies are used to investigate poultry behaviour.

2.2.2. Animal-worn Equipment: 3D-ArUco marker and vest

Selection of ArUco Markers for Detection in Poultry Barn. We selected the ArUco DICT_4X4_1000 dictionary because it stood out compared to others (DICT_5X5, DICT_6X6, DICT_7X7) that we had tested beforehand with chickens. This dictionary offers 1000 distinct patterns, minimizing the risk of overlap in marker identification. The markers were easily detectable across a range of image resolutions, even when tracking fast-moving chickens running from the barn to the winter garden. Each ArUco marker consists of a grid of 4x4 squares, resulting in 16 small squares, or bits, which form a unique binary pattern. Additionally, it is open-source through OpenCV and accessible at https://chev.me/arucogen/. For our study, we used the first 22 markers from this dictionary, excluding marker number 17 (see Appendix A) due to its tendency to cause frequent false detections during preliminary tests in our poultry barn. Marker 17 was too simplistic, and at times, the poultry barn features shared visual similarities with that marker's pattern, creating visual noise for the detection algorithm and leading to false detections. This was particularly the case with the plastic slats with hollow white edges, which were sometimes mistaken for the pattern of marker 17.

Dimensions of ArUco Markers. We maintained a white buffer zone around the markers. This zone, measuring 0.2 cm on all sides, framed the unique pattern of the marker, making it easier to distinguish for automated detection. A sufficient buffer space ensured that the ArUco pattern stood out clearly and was readily detectable, minimising environmental interference or potential stains that could mislead the algorithm. Our markers, measuring 3.4 cm square (1:1 square aspect ratio), were printed on white paper before being laminated to ensure durability. With the buffer zone included, each marker covered a total of 3.8 cm per side. As for the size chosen, the challenge was to find a balance: large enough to be detected by the video processing algorithm but small enough not to disrupt the animal's natural behaviours. The size was also determined by the available space on the chicken's dorsal region—the area between the neck and the tail—where the vest could hold the marker securely without impairing movement. Several tests were carried out beforehand with various sizes to arrive at this effective compromise, ensuring detection while considering the anatomical constraints of the bird.

3D-optimised Design of the ArUco Markers. Although the flat 2D markers could be practical, they may have limitations in terms of the angle of detection. Therefore, we innovated with a 3D design, a cube-like marker (Fig. 2). This 3D-marker, with each face measuring 3.8 cm per side and the entire marker weighing 4.5 g, enhances detection by featuring the same ArUco pattern on each face. The cubic design would ensure that at least one marker was visible, regardless of the animal's orientation, covering almost every position and direction. It was made from durable laminated paper glued to a lightweight foam support. The effectiveness of this design is backed up by previous research from Eagan et al. (2022) and Vagvolgyi et al. (2022), suggesting that placing several copies of the same ArUco pattern on the target can maximise visibility and detection, thus ensuring continuous identification of the animal, regardless of its orientation.

Chicken Vests: Supports for both Automated Detection and Manual Identification. Specially designed chicken vests, High-Vis Chicken Jacket® (Omlet, UK), were used for two main functions. The first was the carriage of the ArUco 3D marker, strategically positioned in the centre of the vest, between the wings and at a roughly equal distance from the chicken's neck and tail. After preliminary testing, this location ensured optimum marker visibility during recording. The second use of these vests was as the primary means of identification, with a unique number written manually on each vest. This identifier, written with indelible ink, could be easily seen by the annotators when viewing the videos, allowing independent validation of the results of the automatic tracking system (Fig. 2). For the 21 hens, vests in three different colours — green, blue, and purple — were used. This colour diversity facilitated annotation, with each animal easily distinguishable, particularly when the vest's colour was considered in addition to the number written on the marker (Appendix A).

During data collection, the vests were worn discontinuously, removed after a maximum of 24 h, and then put back on the following day. Each hen was quickly fitted with a vest held in place by Velcro fasteners, ensuring a comfortable fit. The vests were mainly worn on days when the temperature was below 20 degrees Celsius. When the animals were fitted with the vests, IDs were randomly assigned to individuals, ensuring that the same subjects did not wear the same ID throughout the study; this approach ensured that the vest reading performance was not affected by the animal-specific routine.

2.2.3. Camera configuration and installation

Seven Axis network cameras, 4 of the M1135-E model and 3 of the M1135-E MK II model (Axis Communications, Sweden), were used for this study. With high-definition resolution, these cameras were equipped with a 2-megapixel lens (1920 × 1080) with a horizontal field of

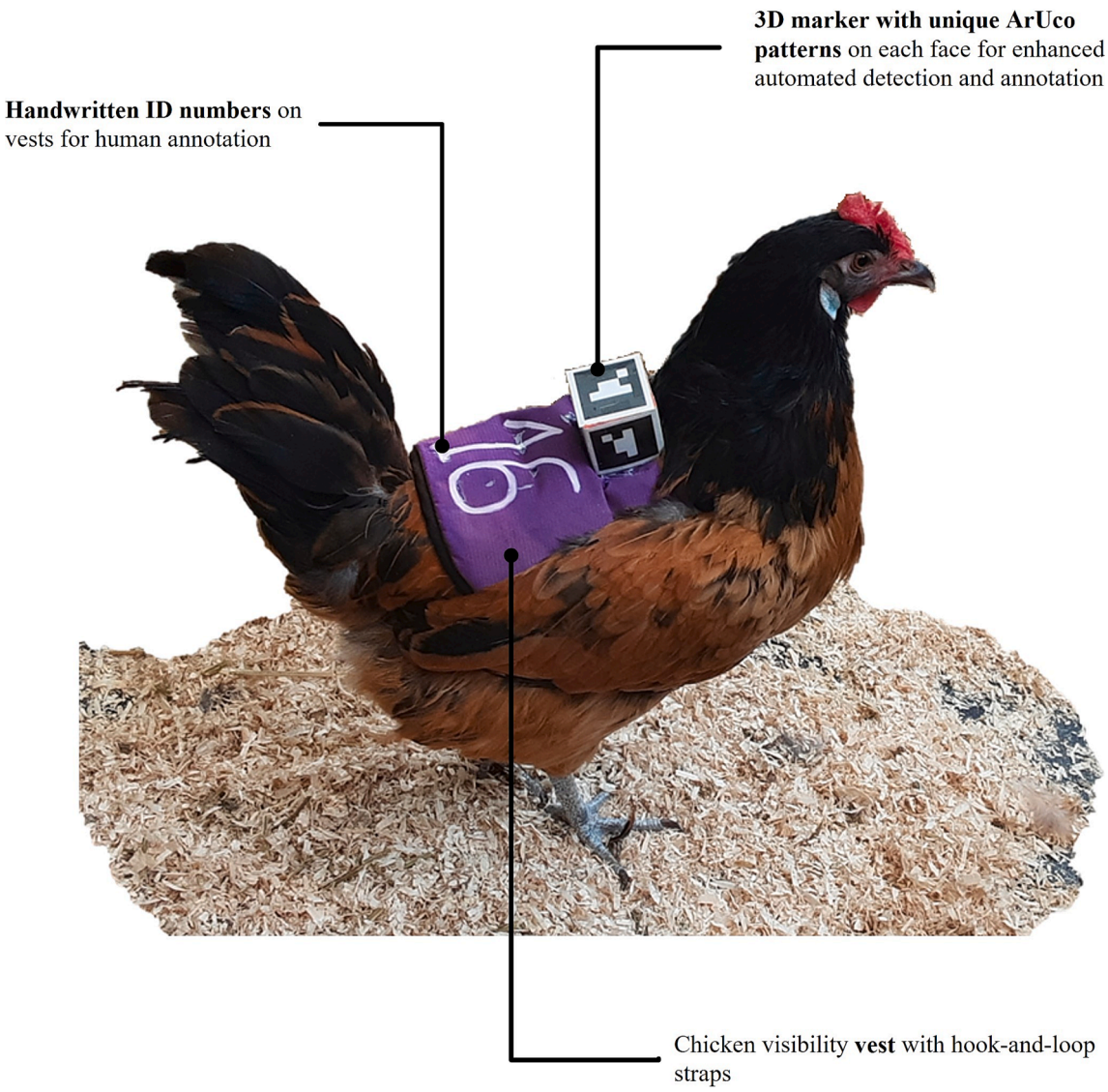


Fig. 2. A chicken from our validation study outfitted with a visibility vest with "16" as its handwritten ID number. On its vest, there is also a cube fitted with the pattern associated with ID 16 from the ArUco_4x4_1000 library, with that pattern displayed on all five visible sides of the cube.

view of 92 degrees. The lens's focal length varied from 3.0 mm to 10.5 mm, ensuring wide coverage and vision of the area of interest. Both camera models had a light sensitivity of 0.1 lx. AXIS Camera Station software was used to set up the cameras.

All settings, from camera position to focal length, were adjusted before recording to ensure optimum image quality. To optimise the view and minimise blind spots, we fixed the cameras to the barn ceiling using suitable clamps, ensuring easy rotation thanks to an adjustable mount with a pivot range of up to 180 degrees. Due to the physical constraints of the environment, overhead positioning was not always possible; most of the cameras were tilted slightly to adjust the viewing angle. Each camera was positioned at a specific distance from its target area, at a determined height above the slatted floor or litter, and oriented at a particular angle relative to the vertical axis. Details of the camera model and position measurements are presented in Table 1.

The video was recorded using the H.264 codec in the MP4 format. We opted for high frame rates to capture the chickens' rapid movements precisely. To put this into perspective, 30 frames per second video means we capture 30 successive pictures of a chicken moving through space, all within one second.

The camera system utilised Power over Ethernet (PoE) technology, which means they are powered directly through the ethernet cable. All

Table 1
Characteristics and positional measurements of cameras at key resources.

No.	Location of camera	Model	Distance	Angle	Frame rate
1	Drinker	M1135	1.18 m	35.86°	25f/s
2	Enrichment facilities	M1135	2.09 m	32.41°	30f/s
3	Feeder	M1135-E MK II	1.61 m	34.06°	30f/s
4	Indoor litter area	M1135-E MK II	1.95–2.50 m	19.34°	30f/s
5	Metal perch	M1135	1 m	49.46°	25f/s
6	Pophole	M1135-E MK II	2.277 m	31.57°	30f/s
7	Wooden perch	M1135	0.77–1.35 m	Overhead	30f/s

the cameras were interconnected via an ethernet network and linked to a central workstation computer.

2.2.4. Workstation computer configuration

For our study, we used a Lenovo ThinkStation P360 Tower computer. This is equipped with a 12th generation Intel® Core™ i7-12700 processor running at 2.10 GHz, a 16.0 GB DDR5 UDIMM memory (15.7 GB usable), an NVIDIA T600 video card with 4 GB video memory, a 512 GB M.2 SSD, and two hard disks with a capacity of 4 TB each.

2.3. Data acquisition and video sequence preparation

2.3.1. Video recording and sequence export

The recordings spanned five days and were conducted only when the animals were fitted with the body-worn equipment (Fig. 2). The video recordings were set to capture automatically, covering only the illuminated periods in the barn, typically from 5:30 am to 9:30 pm. All raw videos collected pertained to the seven resources of interest, capturing both the presence and absence of animals near these resources.

After collecting the raw videos, we split them post-recording into 10-minute sequences using Axis software. This splitting was done to simplify manual annotations and reduce the need to process long video stretches for automated analyses. The total number of manually segmented videos per resource was as follows: metal perch (45), pophole (45), drinker (72), wooden perch (72), feeder (80), enrichment (80), and indoor litter area (80). For each of the seven key resources, 35 sequences were randomly selected from the total segmented sequences, ensuring representativeness across the five-day recording period involving all 21 chickens. The selection process ensured that sequences were distributed across different times of the day—early morning, midday, and evening—based on their timestamps to account for varying lighting conditions.

2.3.2. Pre-processing of video sequences

Our video sequence pre-processing primarily involved delineating a Region of Interest (ROI), marked on each video to identify where most interactions between animals and a specific resource occurred. This ROI, essential for accurate manual and automated annotations, encompassed the resource's main functional area. We used Python's OpenCV library (v4.8.0) to define these ROIs, drawing boundary lines around the resource's functional zones. Defining the ROIs involved three steps (Fig. 3; Appendix B).

Starting with a static camera view focused on the resource, we used this consistent visual reference to extract images from our recordings, showing actual animal interactions with the resource. One randomly selected image served as a baseline for outlining the resource's functional area. This process involved manually marking specific points on each image, such as the tip of a chicken's tail or the position of a 3D-marker on chickens genuinely using the resource. These markings, capturing the primary area where animals interacted with each resource, considered behaviour-specific traits (Table 2). We then converted these points into coordinates, overlaying them onto our baseline image to sketch the ROI.

In the final step, the defined ROI from the baseline image was highlighted while deliberately blurring the peripheral regions. The delineated ROI was then superimposed onto each frame of the videos, producing videos where only the animals within the ROI were distinctly visible.

2.4. Manual annotation using BORIS: Reference or gold standard for validation

We employed human observation as a benchmark to evaluate the effectiveness of the proposed automated annotation technique in accurately identifying animal presence within regions of interest (ROIs). Two observers manually annotated all the pre-processed video sequences using BORIS (v.8.21.5).

Following a pre-established annotation protocol, the annotations were performed on pre-processed video sequences in which only the ROI was visible. The manual annotation of video sequences was based on the mere appearance within the ROI. It commenced when the midsection of the animal's back crossed into or out of the ROI's limits. In cases where the first criterion is not applicable, it started or ceased when both feet were fully positioned inside or outside the ROI, respectively. Here, the

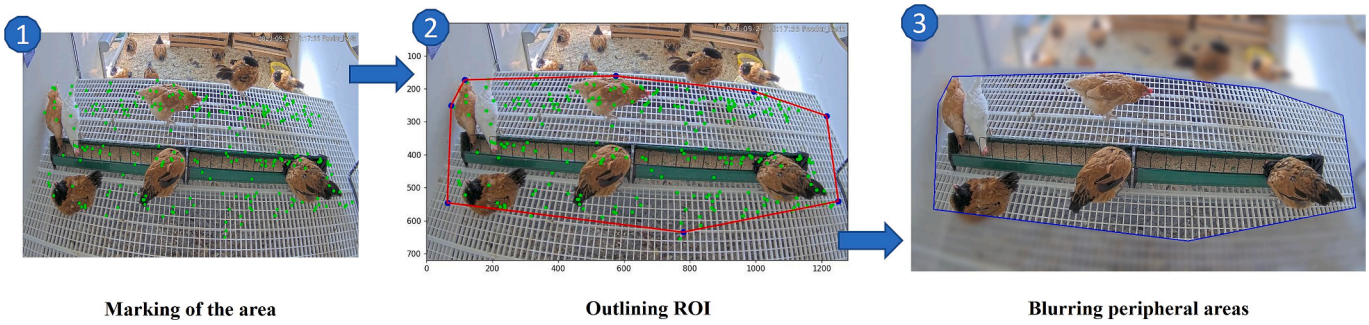


Fig. 3. Definition of Regions of Interest (ROIs), using the feeder as an example. In Step 1, the areas primarily occupied by animals using the resources are marked. In Step 2, an ROI that encompasses most of the marked area is manually outlined. In Step 3, the defined ROI is highlighted, and the peripheral regions are intentionally blurred.

Table 2
Behaviour-specific characteristics used for defining the various regions of interest.

Behaviours	Indicators of effective resource use
Feeding	The chicken is in a lowered head position with the beak in the feeder, pecking at the feed.
Drinking	The chicken is observed standing in front of the drinking line, either with its beak inside a drinking cup, at a nipple, or with its head raised, utilising gravity to ingest the water.
Perching	The chicken is observed sitting or standing on a perch line, using both feet.
Use of the indoor litter area	The chicken is observed standing, walking, scratching the ground, or lying down in the litter area of the barn.
Use of barn-to-winter garden pophole	The chicken passed through the access pophole, leaving the main barn compartment for the winter garden or returning to the barn.
Use of enrichment facilities	The chicken is observed near enrichment elements like the pecking stone, oyster shells, and alfalfa bales, either pecking directly at them or showing interest in nearby fragments on the floor.

specific behaviour of the animal (such as whether it was effectively using the resource or not) was not a determining factor. This ethogram is applied similarly across all resources. The annotators used predefined keys to start and stop the annotations. Handwritten IDs, visible on the videos, were used to track the subjects. Where necessary, the colour of the vest worn by the chickens was also considered to ensure correct manual annotations.

Measures were taken to guarantee reliability and inter-rater agreement. Before starting the actual annotation, a preliminary training phase was conducted, during which the two annotators were trained by an experienced researcher. This training involved coding sample video sequences and discussing the annotation protocol in detail. A sample of 27 video sequences, comprising 3 to 4 clips for each of the seven focus resources, was used for this preliminary phase. The aim was threefold: (1) to identify behavioural coding ambiguities using BORIS, (2) to practise scoring, and (3) to ensure excellent inter-rater reliability between the two annotators. The annotators agreed on the rules for annotation and practised coding before beginning the actual 245 video sequences. Data recorded via BORIS included the subjects' ID, start time, end time, and duration of presence within the ROIs.

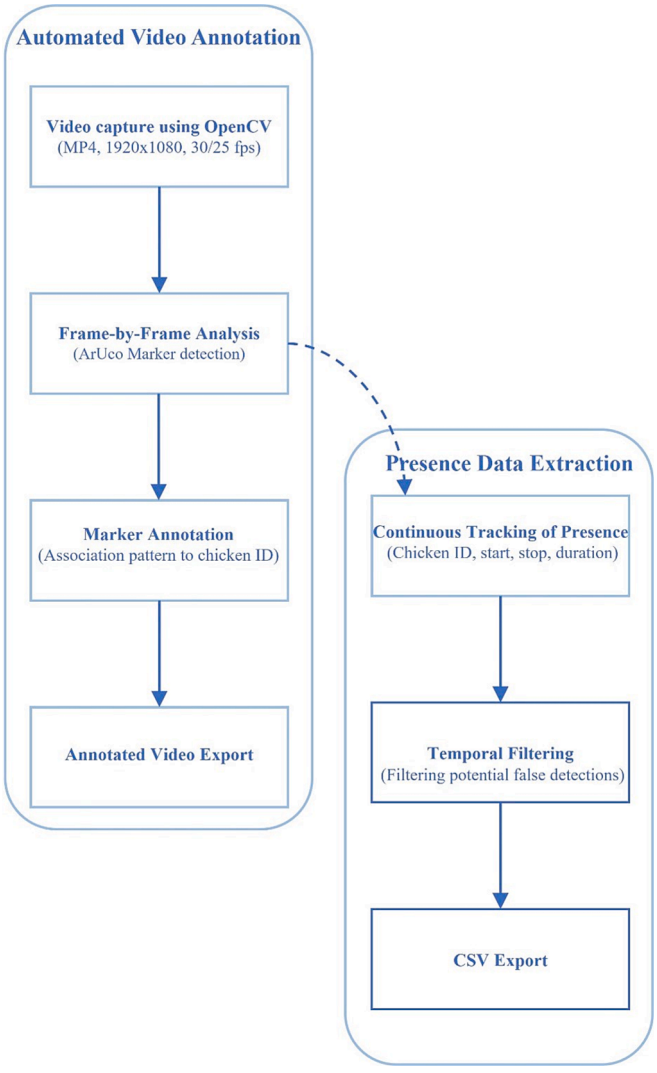


Fig. 4. Flowchart outlining the computational framework used for automated annotation of animal presence within Regions of Interest. Outputs included annotated videos displaying individuals annotated with their respective IDs and the CSV containing data on the presence of the detected individuals.

2.5. Marker detection and automated annotation

The "cv2.ArUco" module of the OpenCV (v4.8.0), a Python library for automated video processing and computer vision (OpenCV Developers, 2023), was employed. This library detected the ArUco markers. Pandas v2.0.3 (The pandas development team, 2023) was used to organise and export the data. Our computational framework primarily consisted of two steps: (1) video processing, which produced a video annotated with chicken ID, and (2) data extraction, resulting in a CSV file that listed the appearances of the various IDs (i.e., individual chickens) over time (Fig. 4).

The first step was to open the pre-processed video sequences with OpenCV's "VideoCapture". These videos, all in MP4 format with a 1920 × 1080 resolution and frame rates of 30 or 25 frames per second, and already pre-processed to highlight only the regions of interest, were our input. Each video was then analysed frame-by-frame. At each frame, the presence of ArUco markers was checked, and once detected, each marker pattern was associated with its predefined unique ID. We adapted the algorithm to filter relevant markers: only markers between the first 22 of the ArUco.DICT_4X4_1000 dictionary, except 17, should be read. Each detected marker was colour-framed to highlight it, and an ID, such as "Chicken 16", was displayed alongside it. These annotated images were then compiled together to create the annotated video.

The second step, data extraction, was also based on frame-by-frame analysis. The extracted data involved the duration of each chicken visit. We inferred the presence of specific chickens at each instant using the known frame rate of the input videos and the number of successive video frames in which individuals appeared. For example, in a video at 30 frames per second, detecting an ID in one frame signifies a presence duration of 1/30th of a second, i.e., approximately 33.33 ms. Therefore, continuous presence in 100 successive frames equate to a total duration of presence of 3.33 s. The visit duration was determined by accumulating moments of presence across successive video frames. A visit was considered ongoing until there was an interruption—loss of continuous detection, either due to the chicken leaving the ROI or the marker becoming obstructed or undetected—exceeding one second without detection. At that point, we considered the visit event as interrupted and recorded a new visit event if the same chicken appeared again in the video sequence.

The data extracted was filtered before being exported as a final CSV output. ArUco marker detection can lead to false readings, sometimes detecting markers from random objects in the environment (Hurník et al., 2021). To address these issues, we employed temporal filtering, particularly because initial tests revealed frequent false readings in certain barn sections. Most errors occurred in specific frames 1, 2, and 32, regardless of frame rate (in fps). Therefore, filters were applied to these specific durations. For example, with a 30-fps camera, we excluded detections of exactly 0.033 s, 0.067 s, and 1.067 s; for 25 fps, 0.04 s, 0.08 s, and 1.28 s were omitted. The cause was likely due to instability in the video encoding process at the time of recording, where suboptimal buffering and encoding settings, particularly in compression and bitrate control, led to detection errors in specific frames. The data, including chicken IDs, start and end times, and duration, were saved in a CSV files for analysis.

2.6. Validation metrics for comparison of automated annotation vs. Human reference

Validation of a tool is the process whereby its measurements are compared with an established reference. This validation usually takes place over a short period and involves calculating various metrics to assess the tool's performance. We employed two distinct approaches to evaluate the performance and validate the proposed automated annotation tool.

The first approach involved a correlation analysis focused on the aggregated presence duration per ID in each video sequence, as detected

by automated annotation, and the corresponding duration observed by human annotators. This approach, which only considered presence data, provides an insight into the agreement between the two annotation methods. We used Spearman's correlation for this analysis, as the aggregated durations were not normally distributed.

The second approach, involving the calculation of classic performance metrics, was inspired by the principles of validation as explained and illustrated by Adrion et al. (2020), Alindekon et al. (2023), and Siegford et al. (2023). In this approach to validation, where both presence and absence could be taken into account, several steps were necessary (Fig. 5). The first step involved chronologically aligning data from the automated annotation with the manually annotated reference, followed by a second-by-second comparison of both methods to determine classification outcomes (also known as confusion matrix elements):

- True Positive (TP): second when automated and manual annotations agreed on the presence of an individual within a specific region of interest.
- True Negative (TN): second when both methods agreed on the absence of an individual within a specific region of interest.
- False Negative (FN): second when the automated system missed an individual that the manual annotation had identified within a specific region of interest.
- False Positive (FP): second when the automated system reported an individual that the manual annotation had not identified.

Next, we standardised TP, FP, TN, and FN sums at the video sequence level, adjusting for varying chicken presence across resources by dividing these counts by the total presence duration per sequence. This standardisation was crucial to account for longer presence yielding more detection instances and to offset the tendency of animals to spend more time away from resources, as Adrion et al. (2020) and Alindekon et al. (2023) noted. Standardisation minimised bias, ensuring a fair comparison of classification outcomes across resources.

Then, for each video sequence, we computed classic validation metrics to evaluate the effectiveness of the proposed automated annotation tool. Metrics were calculated individually for each video sequence using the standardised TP, FP, TN, and FN counts. We considered 35 video sequences per resource, treating each sequence as an individual evaluation unit and as a separate replicate to assess the performance of the corresponding resource. Ultimately, the average of these metrics across units was used to characterise the resources. The performance metrics used were:

- Sensitivity, obtained by the formula: TP_{std} / (TP_{std} + FN_{std}). Sensitivity refers to the proportion of seconds when an ID was present within a given ROI and the system correctly identified it as such.
- Specificity, with the formula: TN_{std} / (TN_{std} + FP_{std}). Specificity assesses the ability of our automated approach to correctly identify seconds when a targeted ID is absent in the ROI.
- Accuracy, which was calculated as: (TP_{std} + TN_{std}) / (TP_{std} + TN_{std} + FP_{std} + FN_{std}). Accuracy determines the overall ability of the

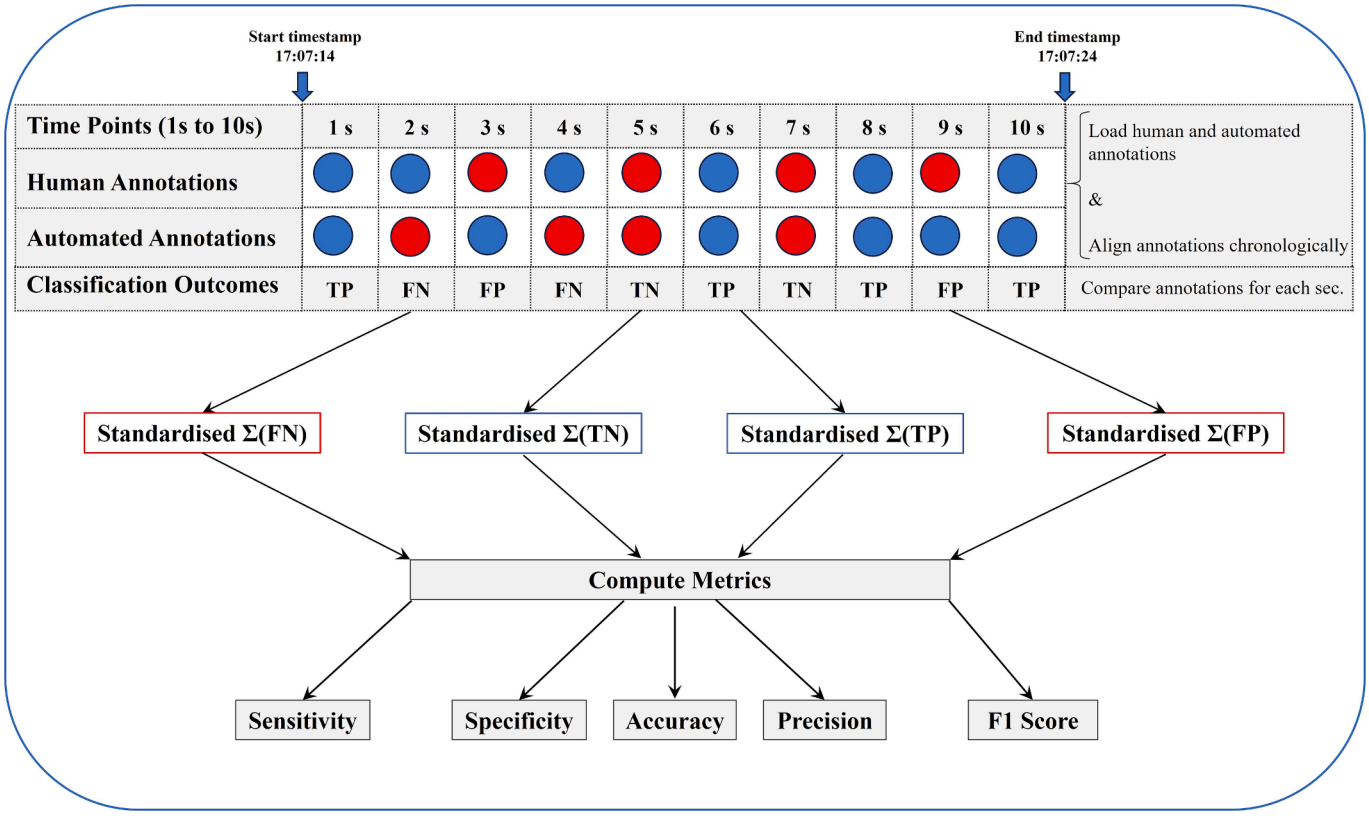


Fig. 5. Diagram summarising the procedure used to calculate the classic performance metrics employed to validate automated annotation against a manual reference over the time points that constitute a video sequence. Here, a 10-second-long video sequence is considered for illustration purposes. The comparison process begins with loading both datasets, aligning them chronologically, and comparing their annotations for each video second. Each circle represents the annotation for a specific second, coloured based on a given individual's presence (blue) or absence (red). Each comparison results in one of four outcomes (or confusion matrix elements): True Positives (TP), True Negatives (TN), False Positives (FP), or False Negatives (FN). Next, the standardised sums of TP, FP, TN, and FN that would account for the total duration of the chickens' presence in each video were calculated. In this illustration, standardised sums of TP, FP, TN, and FN can be respectively obtained as follows: TP_{std} = 4/6, FP_{std} = 2/6, TN_{std} = 2/6, FN_{std} = 2/6 (with the total duration of presence in the video, obtained from human annotation data, as the denominator). Based on these results, classic validation metrics—such as Sensitivity, Specificity, Accuracy, Precision, and F1-Score (balanced average of precision and sensitivity)—are computed to assess the performance of the automated annotation.

Table 3
Comparison of video annotation durations between our automated system and human annotators for each key resource.

Key Resources	Total Video Length (in hours)	Automated Annotation Duration (in hours)	Manual Annotation Duration (in hours)	Time Savings (%)
Drinker	5.86	5.69	19.17	70.32
Enrichment	5.85	4.69	32.50	85.57
Feeder	5.85	6.46	42.33	84.74
Indoor litter area	5.85	6.64	62.67	89.40
Metal perch	5.85	4.12	18.17	77.33
Pophole	5.85	4.49	32.00	85.97
Wooden perch	5.85	6.41	16.00	59.94

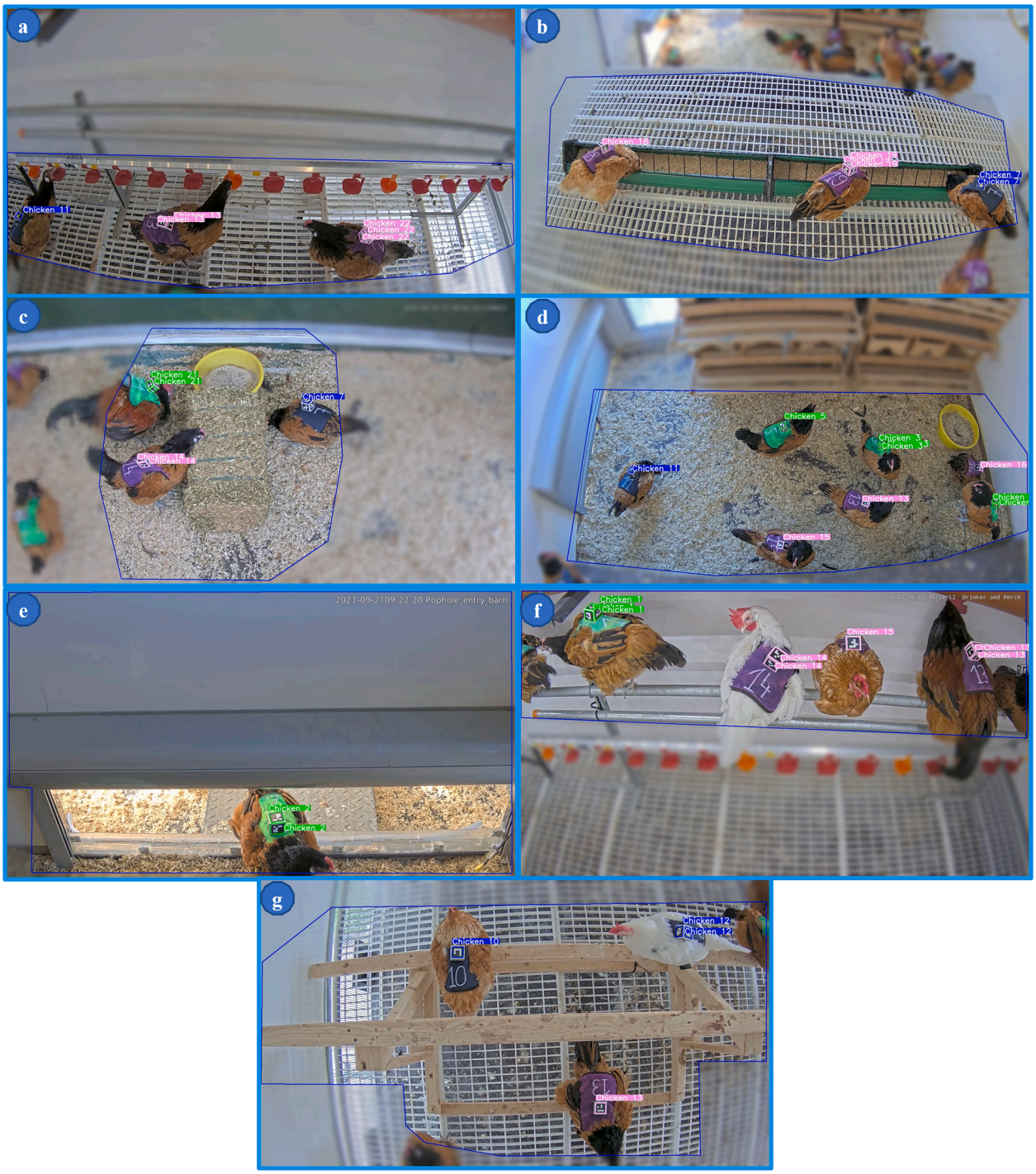


Fig. 6. Selected snapshots from video outputs after automated annotation of ArUco markers on animals present in resource-specific areas of interest—clearly defined, non-blurred zones delineated by thin blue lines. a = region of interest (ROI) around the drinker, b = ROI around the feeder, c = ROI around the enrichment material, d = ROI around the indoor litter area, e = ROI around the pophole, f = ROI around the metal perch, g = ROI around the wooden perch.

automated annotation tool to correctly identify seconds, whether the targeted ID is present or absent.

- Precision, calculated as: TP_{std} / (TP_{std} + FP_{std}). Precision quantifies the proportion of seconds identified as positive (i.e., ID present within ROI) that were actually correct.
- F1-Score, with the formula: 2 × (Precision × Sensitivity) / (Precision + Sensitivity). This metric serves as a harmonic mean that balances Precision and Sensitivity, providing a single score to represent the reliability of the automated annotation tool in identifying chicken presence within the ROI.

Lastly, using Python libraries such as SciPy (v1.11.1) and Scikit-posthocs (v0.8.0), we conducted statistical analyses to explore the influence of various resources on standardised confusion matrix elements (TP_{std}; FP_{std}; TN_{std}; FN_{std}) and derived performance metrics. The Kruskal-Wallis test was employed for these comparisons, as all data distributions exhibited either right or left skewness.

3. Results

We conducted this study to propose an automated method, based on 3D-ArUco markers, to automatically annotate animal presence within regions of interest, namely within functional areas of key resources in poultry farming. A total of 245 10-minute-long sequences of videos of chickens interacting with seven key resources were used. The automatic annotation made it possible to annotate all these video sequences in 40.96 h, an 82.72 % time saving compared to 222.84 h of continuous manual annotation by human observers in front of computer screens (Table 3). Validation of the automated annotation, carried out by two human reference annotators, with an index of inter-rater reliability—Cohen's Kappa, $k = 0.95$ (based on seven 10-min videos)—led to the establishment of performance metrics.

3.1. Output from the developed program

The developed program enabled automated annotation of individuals, generating videos in which individuals' IDs were displayed alongside markers for all visible sides of the cube as long as they were within the region of interest. It offered behavioural data that quantified presence, indicating start and end timestamps, as well as the duration in seconds of each ID during its presence in the region of interest, and compiled these measurements into a dataset. To illustrate and demonstrate these stated capabilities, we present snapshots of some annotated videos in Fig. 6; short annotated video clips are provided in the Supplementary material (Clip 1 to 7), and Fig. 7 offers a summary of the duration data capturing the presence of markers in the region of interest in selected videos.

3.2. Performance evaluation of the automated system

3.2.1. Correlation automated vs human annotations based on aggregated presence duration

Spearman's correlation analysis on all resources revealed an overall correlation coefficient of $\rho = 0.96$ ($p < 0.01$, $N = 1457$). This was calculated with the aggregated duration of the presence of each ID in the videos, comparing automated annotations with human ones (Fig. 8).

When further broken down by resource, the lowest correlation was observed with videos recorded on metal perch ($\rho = 0.56$), while those for all others ranged from 0.92 to 0.99 (Table 4).

3.2.2. Standardised classification outcomes (or confusion matrix elements) across resources

The standardised classification outcomes varied according to resources (H-test, $p < 0.01$; Table 5). Feeder visit recordings showed the highest true positives, contrasting with enrichment, indoor litter, metal perch and pophole. True negatives peaked in pophole recordings and were lowest in the indoor litter area.

While false positives were generally low, they were highest in videos of wooden perch, drinker, and feeder— resources for which plastic slats appeared in the regions of interest. Conversely, indoor litter, metal perch, and enrichment videos had the fewest. Pophole videos presented a mix of these trends. False negatives were most frequent in metal perch, indoor litter, enrichment, and pophole videos, with metal perch notably higher in raw values. Feeder videos recorded the fewest false negatives.

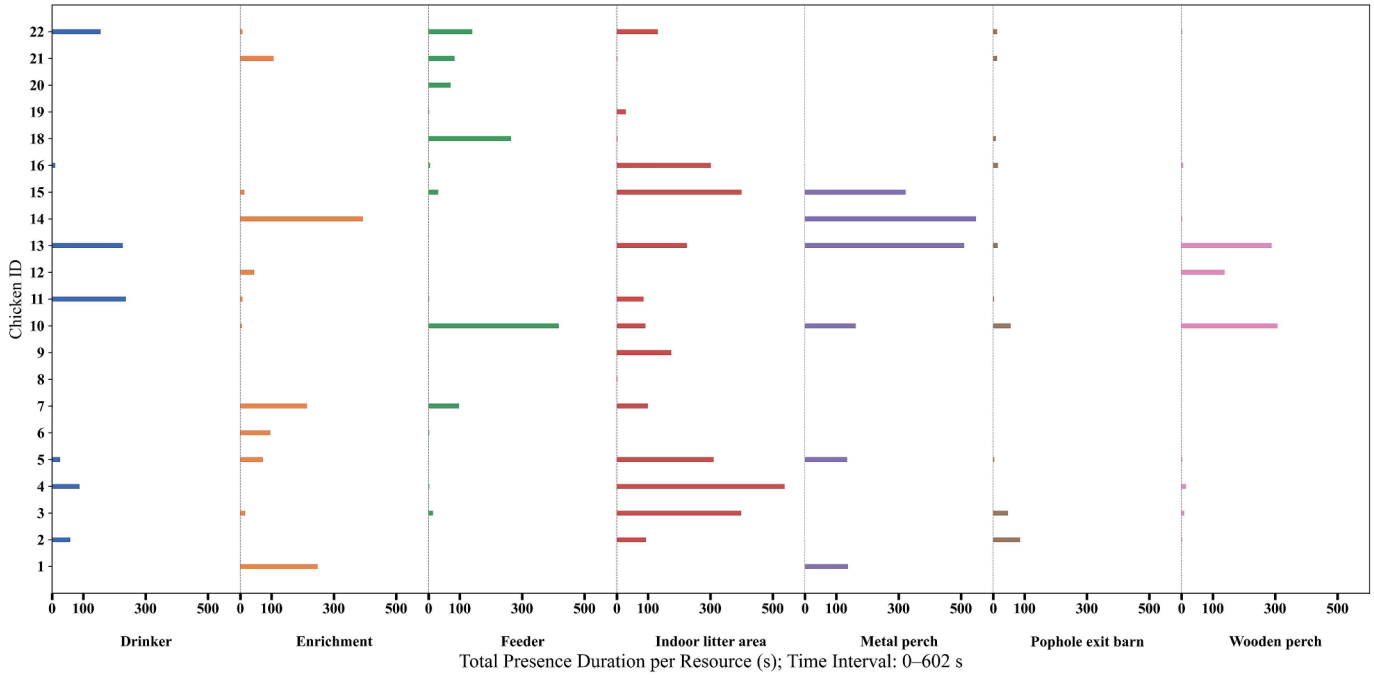


Fig. 7. An illustration of possible behavioural data output (aggregated visit duration for each ID at various resources of interest for a specific timespan). The data were generated along the annotated videos. Video sequences featuring the maximum number of observed individuals per resource were selected for optimal representation.

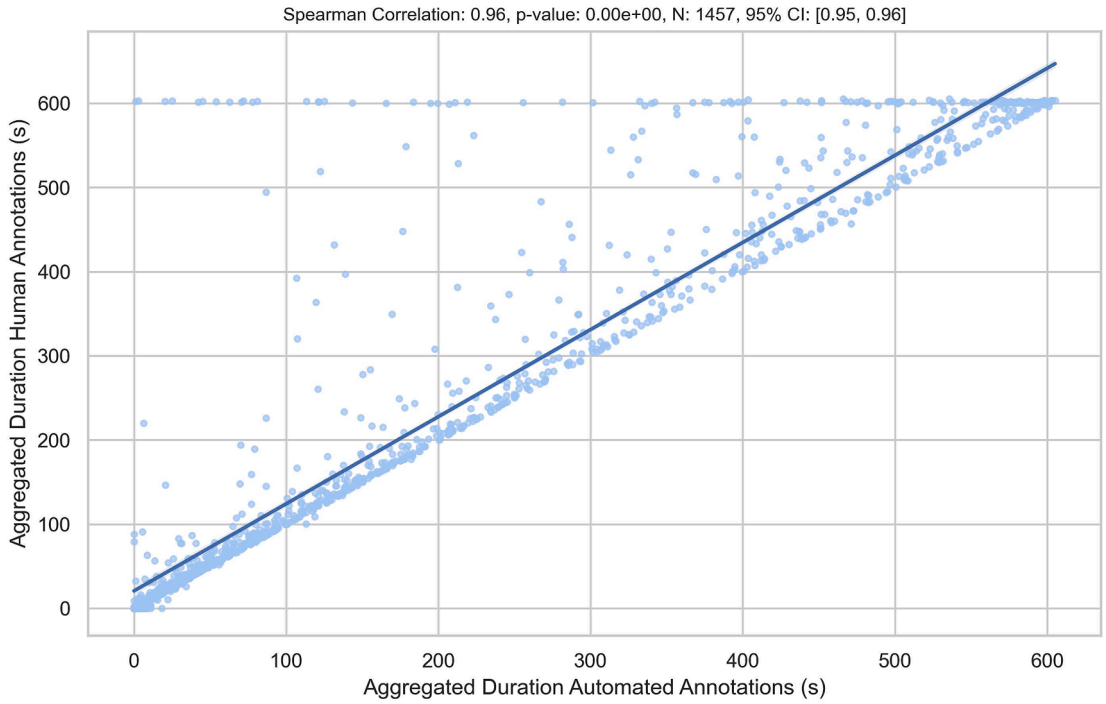


Fig. 8. Scatter plot illustrating the Spearman correlation between automated and human annotations. Light blue data points represent aggregated presence duration for each ID per video.

Table 4
Spearman correlation coefficients by resource based on aggregated presence duration per ID for comparing automated and human annotations.

Key Resources	Spearman's rho (r)	95 % confidence interval	p-value	N
Drinker	0.97	[0.96, 0.98]	< 0.01	141
Enrichment	0.98	[0.98, 0.99]	< 0.01	202
Feeder	0.98	[0.98, 0.99]	< 0.01	254
Indoor litter area	0.92	[0.90, 0.93]	< 0.01	380
Metal perch	0.56	[0.44, 0.69]	< 0.01	117
Pophole	0.99	[0.98, 0.99]	< 0.01	195
Wooden perch	0.94	[0.92, 0.96]	< 0.01	168

Examples of false positive and false negative cases are provided in Appendices C and D, respectively.

3.2.3. *Metrics-Based performance overview*

The automated annotation method showed overall performance metrics of 92.83 % sensitivity, 99.93 % specificity, 99.08 % accuracy, 98.77 % precision, and a 95.28 % F1-Score, with significant variations across resources (H-test, $p < 0.01$; Table 6). Feeder visit recordings had the highest sensitivity, in contrast to the enrichment, indoor litter area, metal perch and pophole, with the metal perch showing notably lower sensitivity at 81.29 %. A similar trend was seen in F1-scores.

Specificity was generally high, exceeding 99 % in all resources, with the lowest figures at the wooden perch and feeder; the differences stemmed from the decimals. The lowest accuracies, around 98 %, were recorded in metal perch and indoor litter area, while other resources exceeded 99 %. Precision values were lowest for wooden perch, pophole, drinker and feeder, ranging from 98 % to 99 %, compared to above 99 % for other resources.

Table 5
Comparison of standardised values of the classification outcomes: True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN), accounting for total durations of chicken presence in each video.

Key Resources	TP_{std}	TN_{std}	FP_{std}	FN_{std}
Drinker	0.95^{ab}	27.15^{b}	0.02^{a}	0.05^{ab}
Enrichment	0.95^{b}	24.84^{b}	0.00^{b}	0.05^{a}
Feeder	0.98^{a}	10.37^{b}	0.01^{a}	0.02^{c}
Indoor litter area	0.93^{b}	2.70^{c}	0.00^{b}	0.07^{a}
Metal perch	0.81^{b}	10.30^{b}	0.00^{b}	0.19^{a}
Pophole	0.93^{b}	108.11^{b}	0.02^{ab}	0.07^{a}
Wooden perch	0.96^{ab}	15.96^{b}	0.02^{a}	0.04^{ab}

In the table, 'a', 'b', and 'c' categorise groups by statistical significance, with 'a' indicating the highest group. Across columns, values with different superscripts are statistically different ($p < 0.05$).

Table 6
Performance metrics for evaluation of validity across resources.

Key resources	Sensitivity (%)	Specificity (%)	Accuracy (%)	Precision (%)	F1-Score (%)
Drinker	94.80^{ab}	99.92^{ab}	99.75^{a}	97.98^{b}	95.96^{ab}
Enrichment	94.55^{b}	99.97^{a}	99.60^{ab}	99.59^{a}	96.96^{b}
Feeder	97.63^{a}	99.89^{b}	99.70^{ab}	98.82^{b}	98.09^{a}
Indoor litter area	92.78^{b}	99.93^{ab}	97.76^{c}	99.82^{a}	96.10^{b}
Metal perch	81.29^{b}	99.97^{ab}	97.36^{bc}	99.62^{a}	88.32^{b}
Pophole	92.79^{b}	99.97^{ab}	99.71^{ab}	97.78^{b}	94.72^{b}
Wooden perch	95.95^{ab}	99.87^{b}	99.65^{ab}	97.77^{b}	96.79^{ab}

In the table, 'a', 'b', and 'c' categorise groups by statistical significance, with 'a' indicating the highest group. Across columns, values with different superscripts are statistically different ($p < 0.05$).

4. Discussion

With the aim of promoting automation in the annotation of animal presence within a region of interest, which is crucial for validating and adopting new technologies in ethology, our paper proposed an automated annotation method based on 3D-ArUco markers attached to the

animals' backs. This method has been detailed and can be used to acquire annotated videos showing individuals' IDs as they move through a delimited region of interest (ROI), thought to represent the functional area of the resources used. It can also collect individual-focused behavioural data, including the start and end timestamps of each presence within the ROIs. The validation performance of this automatic behavioural annotation tool, compared with human observations, the limitations of the tool, and suggestions for its optimal deployment are discussed.

4.1. An interpretation of the system performance through multiple indicators

In our study, we used six key indicators to evaluate the quality and suitability of the proposed automated annotation method. This multidimensional evaluation is essential as it provides a more holistic and reliable assessment of the method's validity. Comparing the outcomes of our study with those of previous studies was challenging, as many authors do not consistently apply classic performance validation indicators. Nevertheless, we have attempted to draw analogies, which may be debatable, based on the calculation approach or aim of the metrics, between our performance measurements and those of other studies.

4.1.1. Spearman correlation

Comparison of the automated annotation against that carried out by human observers revealed a Spearman correlation of 0.96 for the total duration of presence per individual from each video. This value indicates a robust correlation (Martin and Bateson, 2007). Spearman's correlation measures the strength and direction of the relationship between two variables; it provides insights into the overall agreement between the two annotation methods. In practical terms, the observed correlation suggests that for each minute of video, the estimation of presence by the automated system closely matched the human annotations for at least 57 s, underlining its potential as a reliable tracker.

In comparison with the existing literature, our overall correlation analysis, aimed to assess the general agreement between two annotation methods, is in line with the approach, at least in terms of its purpose, of Intraclass Correlation Coefficients (ICC) used by Eagan et al. (2022). Our results corroborate with theirs (ICC of 80 % and 96 %) for validating ArUco markers attached to cats for tracking within regions of interest. Such corroboration is not surprising given our similar methodologies; we both used 4x4-bit markers and placed several unique markers on the same subject to improve detectability. Our method involved multiple markers on the sides of a cube, whereas theirs used multiple unique markers along necklaces. In addition, both studies operated with a high frame rate for video recording, approximately 30 frames per second. The importance of achieving high detection performance becomes evident when considering the use of a substantial number of frames and multiple unique markers on a subject.

4.1.2. Sensitivity

In our study, a sensitivity of 92.83 % was obtained, indicating high performance (Brown-Brandl et al., 2019). This metric evaluates the system's ability to identify actual presence correctly. Such a value implies that, during a one-minute period when tagged animals are within a region of interest, the automated tool would correctly detect them for at least 55 s, demonstrating its strong effectiveness in capturing actual presences within the region of interest.

Compared to previous literature, our sensitivity is higher than the 64.58 % detection rate reported by Alarcón-Nieto et al. (2018) after interpolation for ArUco-like markers attached to domesticated zebra finches. Aside from their subjects being smaller than ours, the discrepancies between our results may be due to our different approaches to mounting the markers. We used a 3D-marker design featuring multiple unique IDs with low potential for physical interaction, minimising marker damage and obscuration, thus improving the chances of detection. The larger size of our markers (38 mm versus 10 mm for theirs) and perhaps the technical capabilities of the specific algorithms and marker dictionaries used, which these authors have not documented, could be other factors contributing to the divergence of results.

4.1.3. Specificity

Our validation study revealed a specificity of 99.93 %, reflecting a robust identification of true negatives: the automatic annotation method showed a significant ability to identify exactly what it is supposed to, without going beyond this objective (Martin and Bateson, 2007; Brown-Brandl et al., 2019). Specificity assesses the validity of a tool in recognising actual absences, which translates into approximately 59 s of accurate non-detection per minute. We have not been able to establish a link with previous literature related to ArUco-like tracking validation using specificity.

4.1.4. Accuracy

An accuracy of 99.08 % was achieved, a value considered excellent (Martin and Bateson, 2007). Accuracy indicates correct classifications of presence and absence, demonstrating the extent to which the system resists systematic error. In practice, the value obtained is equivalent to a margin of error of less than one second per minute, suggesting an exceptional confidence level for annotating tagged animals within a region of interest. We have also been unable to draw parallels with previous literature on ArUco-like tracking for this metric.

4.1.5. Precision

The proposed annotation method had a precision of 98.77 %, indicative of the system's high reliability in true positive detections (Martin and Bateson, 2007; Brown-Brandl et al., 2019). Such a value suggests only 1 s of false detection per minute of a tagged animal within a region of interest.

Compared to the literature, our overall precision is higher than that of Sclocco et al. (2021), who reported a precision of 38.53 % for ArUco tracking in ants. Precision reflects the proportion of true results among all cases examined. Technical differences at the hardware level might explain this variance. They used a lower frame rate than ours (10 versus 25/30 frames per second that we used). A higher frame rate could reduce motion blur, allowing a smoother and more continuous capture of subject movement and adapting more quickly to changes in speed and direction (Sclocco et al., 2021). Additionally, variations in the setup, among others, such as the characteristics of the cameras and the distance ratio between the camera and the surface area of the markers, probably contributed to this observed difference in performance. Moreover, their system relies heavily on a top-down configuration, while our introduction of a cubic marker offers increased visibility in the camera's field of view.

4.1.6. F1-score

Our automated annotation yielded an F1-score of 95.28 %, implying that the system correctly detected presences and excluded non-presences. F1-score represents a harmonious balance between precision and sensitivity. Such a value means that out of 100 detections, 95 are expected to be correctly identified, attesting to the reliability of the proposed automated annotation method. No parallel could be established with previous literature related to ArUco-like tracking using this metric.

In sum, the robustness and reliability of our system can be confirmed given the high performance (mostly $>$ 95 %) demonstrated by the multiple indicators in our analysis. Each of these indicators, contributing its unique perspective, converges to validate the system's ability to annotate animal presence within a region of interest with high reliability.

4.2. Misreadings

Although not significantly affecting overall performance indicators, we observed that in our study, about 5 % of the performance decreases in our annotation system were attributed to misreadings. These errors can be categorised into two groups, namely reading failures (i.e., False Negatives) and false readings (i.e., False Positives), each having distinct causes. Some suggestions and practical arrangements to minimise misreadings when implementing 3D-ArUco markers for annotating animal presence within a region of interest are formulated.

4.2.1. Reading failures (i.e., false negatives)

For ArUco detection to function correctly, the black and white grids composing the marker and the surrounding white buffer zones must be completely visible and free of any occlusions to ensure precise decoding. Visibility issues with ArUco markers as causes of reading failures have often been mentioned in previous studies (e.g., Alarcón-Nieto et al., 2018; Eagan et al., 2022; Wolf et al., 2023). In our study, the highest rate of reading failures was observed with video recordings about the metal perch, followed by the indoor litter area, enrichment and pophole. Three major reasons for reading failures can be highlighted.

Firstly, except for the metal perch, occasional blurriness was observed, particularly in the focus region around most of these resources, where the camera distances were the greatest, ranging from 2 to 2.5 m, compromising image clarity. For the metal perch, there was notably poor camera arrangement with an oblique angle of nearly 50 degrees, limiting visibility from above while allowing only frontal visibility of one side of the cubic marker when the chickens were perching.

Secondly, physical interferences also led to false negatives. These reading failures occurred when chickens adopted stationary positions where the marker became partially or completely obstructed. These obstructions occurred when the birds' feathers masked the marker, e.g., during (night-) daytime perching or preening; they also happened when any other housing component or animal obstructed the camera's view.

A final factor contributing to visibility issues could be related to variable lighting conditions and or direct exposure of the markers, causing reflections that hindered the precise detection of the pattern during processing (Alarcón-Nieto et al., 2018; Crall et al., 2015; Sclocco et al., 2021). We experienced such cases, which are problematic when the animal remains stationary, especially in the indoor litter, pophole or enrichment areas, which are the most exposed to daylight.

4.2.2. False readings (i.e., false positives)

The false readings, referred to as false positives, are misreadings that can result from confusion between the random patterns in the environment and those of the markers (Hurník et al., 2021). In our study, we observed such erroneous readings mainly during the processing of recordings from the drinker, feeder, and wooden perch, and this was anticipated, as these resources were placed on slatted plastic with white edges and hollow patterns that naturally mimic ArUco marker. False readings only happened on rare occasions and only lasted for the duration of a sudden lighting variation, light reflections, glare, motion blur in environment components, and other unidentifiable causes. Despite the implementation of temporal filtering to discard short-lasting false detections, as well as the deliberate selection and reading restriction of the tags on which our developed algorithm should operate, eliminating these errors completely remained seemingly unattainable in those areas of the poultry barn.

4.3. Other limitations of the study

Three limitations, particularly about the methodology, can be noted in this study. One is the loss of the absolute scale of the classification outcomes (i.e., elements of the confusion matrix); these were transformed into ratios relative to the duration of presence. We opted for standardised values rather than absolute counts to allow fair comparison across resources, as chicken presence is over-represented in some resources compared to others. Besides the loss of absolute scale, standardisation assumes a uniform visit of resources (i.e., a homogeneous distribution across all standardised segments) and focuses on average behaviour, assuming a constant level of chicken activity at the resources. Given that the main objective of our study was to validate the use of the ArUco marker and not to focus on the behaviours per se, the assumption of a uniform distribution of resource visits did not affect the achievement of our aim.

The second limitation lies in the risk of inaccurately estimating metrics such as accuracy and specificity. It is possible to observe artificial inflation of these metrics because, in their calculation formulas, negative cases — situations where the chickens are not at the resource — are a determining factor influencing the result, although our standardisation can reduce this influence to a certain extent. Adrion et al. (2020) and Alindekon et al. (2024) highlighted that the number of negatives in animal behaviours is generally very high, with animals spending more time absent than present at the resource. The other performance metrics, such as Precision, Sensitivity, and the F1-Score, are nevertheless less susceptible to such an imbalance in the dataset.

The third limitation of our proposed automated annotation method lies in its inability to distinguish actual resource use from mere presence in the region of interest. Our marker-based tracking system, as we propose it, does not allow us to determine for certain if an animal is genuinely using a resource, but it can confirm with more than 95 % certainty if the animal is in the defined region of interest around the resource. This constraint is similar to that of RFID technology, which also indicates if an animal is in an area of interest but cannot confirm if the animal is actually using the resource (Alindekon et al., 2023). To annotate with certainty genuine resource usage, one must use equipment capable of doing so. For example, for drinker use, a device equipped with water flow meters to measure each animal's water consumption, as in Maselyne et al. (2016); for perch use in chickens, a system that measures the use of perches using a load cell module recording the weight of the animals as in Wang et al. (2019). Advanced technologies, such as those incorporating machine learning algorithms (e.g., Guo et al., 2022; Liu et al., 2023), may also be employed to gain insights into actual resource use.

4.4. Practical insights on applying the 3D-ArUco system for validation

4.4.1. Minimising misreads

Several practical adjustments are required to minimise reading failures and improve the readability of ArUco markers. Instead of laminated paper, non-reflective, 3D-printed supports can be used for better effectiveness. Regular cleaning and maintenance are necessary to ensure optimal performance over time. Camera placement should also be tailored to the behaviours of interest. For static behaviours like resting or sleeping, an overhead camera works best, while in areas with dynamic movements, such as near feeders or drinkers, cameras should be angled at 35 degrees and placed about 1.6 m high.

To prevent false readings, it is crucial to use distinct marker patterns that are not easily confused with the surrounding environment. Conducting a preliminary analysis without markers can help identify potential environmental patterns that might cause algorithmic errors, allowing researchers to choose the most suitable markers for their study.

4.4.2. Cross-validation tool applicable to various technologies

Our 3D-ArUco system can be especially valuable for validating other wearable tracking technologies, provided it does not interfere with their functionality. For example, this could apply to body-worn trackers like RFID, which, like the proposed system, tracks animals and generates spatiotemporal data on their movements. By cross-referencing data from the 3D-ArUco system, researchers can verify the accuracy of RFID outputs.

However, the use of 3D-ArUco markers may present certain

challenges, particularly when validating technologies that integrate vision-based tracking systems. These systems often use object detectors (e.g., bounding boxes; Siriani et al., 2022) to locate individual animals and associate their identities across frames. Problems can arise when two 3D-ArUco markers fall within the same bounding box, potentially leading to identity confusion or tracking inaccuracies. This limitation can also occur in segmentation-based and keypoint-based tracking systems, where overlapping markers or animals may complicate identity association.

4.4.3. *Adapting the 3D-ArUco system to different species and environments*

While this study focuses on poultry, the adaptability of the 3D-ArUco system makes it suitable for various species and environments. Researchers must customise the system based on species-specific factors like marker placement, material durability, and environmental complexity. For example, robust materials are essential for pigs, known for biting and potentially damaging markers. For cattle, which may dislodge vests, using multiple identical ArUco markers on waterproof supports attached to collars, as demonstrated in Sadrzadeh et al. (2024), can be a viable alternative to vests or 3D markers. The successful use of body-worn devices, such as accelerometers and GPS trackers, enduringly fixed to animals like sheep and waterbirds (Kölzsch et al., 2016; Ikurior et al., 2021), suggests that 3D markers could similarly be securely affixed for these species. Since the 3D-ArUco system is intended for short-term validation, it offers flexibility and can be tailored to meet the needs of different species and research objectives effectively.

5. Conclusions

This paper has detailed the implementation and demonstrated the effectiveness of a computer vision method employing cubic ArUco markers for validating animal tracking technologies. The method showed high validation performance in automated annotation of presence within regions of interest and significant time savings for humans, offering a precise and reliable alternative to visual observation. This can reduce observer fatigue, bias and physical discomfort in validating other emerging technologies. Furthermore, the method utilises pre-existing open-source pattern recognition techniques, simplifying its use. This eliminates the need to develop complex new algorithms, making it accessible to a broad range of researchers. This paper demonstrates that introducing reliable technological solutions can support and significantly improve animal behaviour research.

CRediT authorship contribution statement

Serge Alindekon: Writing – original draft, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Jana Deutsch:** Writing – review & editing, Methodology, Investigation. **T. Bas Rodenburg:** Writing – review & editing. **Jan Langbein:** Writing – review & editing. **Birger Puppe:** Writing – review & editing, Resources. **Helen Louton:** Writing – review & editing, Supervision, Resources, Methodology, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors would like to acknowledge the contributions of several persons who greatly contributed to the success of this study and its reporting: Heino Marx, who assisted in customising the mobile barn, setting up equipment, and taking care of the chickens; Klaus-Dieter Witt, who provided the chickens and essential supplies, including food and enrichment materials; Heike-Bettina Riese, who organised and provided animal care throughout the study; Michael Seehaus and Enrico Daum, who assisted with camera and workstation configurations, including the setup of video servers and recording management; Theresa Ludwig, who advised on BORIS software use; Timo Homeier-Bachmann and Anne Schütz, who made part of the equipment available for data collection; and finally, Wiebke Knoblauch and the journal's anonymous reviewers who commented on and edited an early version of the manuscript.

Funding

Serge Alindekon, a doctoral candidate since 2021/2022, received financial support for his research from the German Academic Exchange Service (grant No. 57552340).

Appendices and Supplementary Data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compag.2024.109812.

Data availability

Data will be made available on request.

References

Adrion, F., Keller, M., Bozzolini, G.B., Umstatter, C., 2020. Setup, test and validation of a UHF RFID system for monitoring feeding behaviour of dairy cows. Sensors (switzerland) 20, 1–19. https://doi.org/10.3390/s20247035.

AHAW, 2015. Scientific Opinion on welfare aspects of the use of perches for laying hens. EFSA J. 13. https://doi.org/10.2903/j.efsa.2015.4131.

Alarcón-Nieto, G., Graving, J.M., Klarevas-Irby, J.A., Maldonado-Chaparro, A.A., Mueller, I., Farine, D.R., 2018. An automated barcode tracking system for behavioural studies in birds. Methods Ecol. Evol. 9, 1536–1547. https://doi.org/10.1111/2041-210X.13005.

Alindekon, S., Rodenburg, T.B., Langbein, J., Puppe, B., Wilmsmeier, O., Louton, H., 2023. Setting the stage to tag "n" track: a guideline for implementing, validating and reporting a radio frequency identification system for monitoring resource visit behavior in poultry. Poult. Sci. https://doi.org/10.1016/j.psj.2023.102799.

Alindekon, S., Deutsch, J., Langbein, J., Rodenburg, T.B., Puppe, B., Homeier-Bachmann, T., Louton, H., 2024. Inferring resource use from functional area presence in a small, single-flock of chickens in a mobile barn. Poult. Sci. 103, 104123. https://doi.org/10.1016/j.psj.2024.104123.

Blachut, K., Danilowicz, M., Szolc, H., Wasala, M., Kryjak, T., Komorkiewicz, M., 2022. Automotive perception system evaluation with reference data from a UAV's camera using ArUco markers and DCNN. J. Signal Process. Syst 94, 675–692. https://doi.org/10.1007/s11265-021-01734-3.

Brown-Brandl, T.M., Adrion, F., Maselyne, J., Kapun, A., Hessel, E.F., Saeys, W., Van Nuffel, A., Gallmann, E., 2019. A review of passive radio frequency identification systems for animal monitoring in livestock facilities. Appl. Eng. Agric. 35, 579–591. https://doi.org/10.13031/aea.12928.

Čepon, G., Ocepek, D., Kodrič, M., Demšar, M., Bregar, T., Boltežar, M., 2023. Impact-Pose estimation using ArUco markers in structural dynamics. Exp. Tech. https://doi.org/10.1007/s40799-023-00646-0.

Collias, N.E., Collias, E.C., 1996. Social organisation of a red junglefowl, Gallus gallus, population related to evolution theory. Anim. Behav. 51, 1337–1354. https://doi.org/10.1006/anbe.1996.0137.

Crall, J.D., Gravish, N., Mountcastle, A.M., Combes, S.A., 2015. BEEtag: A low-cost, image-based tracking system for the study of animal behavior and locomotion. PLoS One 10. https://doi.org/10.1371/journal.pone.0136487.

Eagan, B.H., Eagan, B., Protopopova, A., 2022. Behaviour real-time spatial tracking identification (BeRSTID) used for cat behaviour monitoring in an animal shelter. Sci. Rep. 12, 1–9. https://doi.org/10.1038/s41598-022-22167-3.

Ferrão, J., Dias, P., Neves, A.J.R., 2018. Detection of ArUco Markers Using the Quadrilateral Sum Conjuncture. In: Lecture Notes in Computer Science. Springer International Publishing, pp. 363–369. https://doi.org/10.1007/978-3-319-93000-8_41.

Garrido-Jurado, S., Muñoz-Salinas, R., Madrid-Cuevas, F.J., Marín-Jiménez, M.J., 2014. Automatic generation and detection of highly reliable fiducial markers under occlusion. Pattern Recognit. 47, 2280–2292.

Garrido-Jurado, S., Muñoz-Salinas, R., Madrid-Cuevas, F.J., Medina-Carnicer, R., Munoz-Salinas, R., Madrid-Cuevas, F.J., Medina-Carnicer, R., Muñoz-Salinas, R., Madrid-Cuevas, F.J., Medina-Carnicer, R., 2016. Generation of fiducial marker dictionaries using mixed integer linear programming. Pattern Recognit. 51, 481–491. https://doi.org/10.1016/j.patcog.2015.09.023.

Gómez, Y., Berezowski, J., Jorge, Y.A., Gebhardt-Henrich, S.G., Vögeli, S., Stratmann, A., Toscano, M.J., Voelkl, B., 2022. Similarity in temporal movement patterns in laying

hens increases with time and social association. Animals 12. https://doi.org/10.3390/ani12050555.
Guo, Q., Sun, Y., Min, L., van Putten, A., Knol, E., Visser, B., Rodenburg, T., Bolhuis, J., Bijma, P., de With, N.P., 2022. Video-Based Detection and Tracking with Improved Re-Identification Association for Pigs and Laying Hens in Farms. In: Proceedings of the international joint conference on computer vision, imaging and computer graphics theory and applications, pp. 69–78. https://doi.org/10.5220/0010788100003124.
Hurník, J., Zatočilová, A., Paloušek, D., 2021. Circular coded target system for industrial applications. Mach. vis Appl. 32, 2021. https://doi.org/10.1007/s00138-020-01159-1.
Ikurior, S.J., Marquetoux, N., Leu, S.T., Corner-Thomas, R.A., Scott, I., Pomroy, W.E., 2021. What are sheep doing? Tri-axial accelerometer sensor data identify the diel activity pattern of ewe lambs on pasture. Sensors 21, 6816. https://doi.org/10.3390/s21206816.
Koeda, M., Yano, D., Shintaku, N., Onishi, K., Noborio, H., 2018. Development of wireless surgical knife attachment with proximity indicators using ArUco marker. In: Stephanidis, C. (Ed.), Human-Computer Interaction. Interaction in Context: 20th International Conference, HCI International 2018, Las Vegas, NV, USA, July 15–20, 2018, Proceedings, Part II. Lecture Notes in Computer Science, vol. 10903, 14–26. Springer, Cham.
Kölzsch, A., Neefjes, M., Barkway, J., Müskens, G.J., van Langevelde, F., de Boer, W.F., Nolet, B.A., 2016. Neckband or backpack? Differences in tag design and their effects on GPS/accelerometer tracking results in large waterbirds. Anim. Biotelem. 4, 1–14.
Levitis, D.A., Lidicker, W.Z., Freund, G., 2009. Behavioural biologists do not agree on what constitutes behaviour. Anim. Behav. 78, 103–110. https://doi.org/10.1016/j.anbehav.2009.03.018.
Liu, D., Parmiggiani, A., Psota, E., Fitzgerald, R., Norton, T., 2023. Where's your head at? Detecting the orientation and position of pigs with rotated bounding boxes. Comput. Electron. Agric 212, 108099. https://doi.org/10.1016/j.compag.2023.108099.
Martin, P., Bateson, P., 2007. An introductory guide to measuring behaviour, 3rd ed. Cambridge University Press, Cambridge.
Maselyne, J., Adriaens, I., Huybrechts, T., De Ketelaere, B., Millet, S., Vangeyte, J., Van Nuffel, A., Saeys, W., 2016. Measuring the drinking behaviour of individual pigs housed in group using radio frequency identification (RFID). Animal 10, 1557–1566. https://doi.org/10.1017/S1751731115000774.
OpenCV Developers, 2023. ArUco Marker Detection. Retrieved September 7, 2023, from https://docs.opencv.org/3.4/d9/d6a/group_aruco.html.
Othayoth, R., Strebel, B., Han, Y., Francois, E., Li, C., 2022. A terrain treadmill to study animal locomotion through large obstacles. J. Exp. Biol. 225. https://doi.org/10.1242/jeb.243558.
Sadrzadeh, N., Foris, B., Krahn, J., von Keyserlingk, M.A.G., Weary, D.M., 2024. Automated monitoring of brush use in dairy cattle. PLoS One 19, e0305671. https://doi.org/10.1371/journal.pone.0305671.
Sales, G.T., Green, A.R., Gates, R.S., Brown-Brandl, T.M., Eigenberg, R.A., 2015. Quantifying detection performance of a passive low-frequency RFID system in an environmental preference chamber for laying hens. Comput. Electron. Agric 114, 261–268. https://doi.org/10.1016/j.compag.2015.03.008.
Sclocco, A., Ong, S.J.Y., Pyay Aung, S.Y., Teseo, S., 2021. Integrating real-time data analysis into automatic tracking of social insects. R Soc Open Sci 8. https://doi.org/10.1098/rsos.202033.
Siegford, J.M., Steibel, J.P., Han, J., Benjamin, M., Brown-Brandl, T., Dórea, J.R.R., Morris, D., Norton, T., Psota, E., Rosa, G.J.M., 2023. The quest to develop automated systems for monitoring animal behavior. Appl. Anim. Behav. Sci. 265. https://doi.org/10.1016/j.applanim.2023.106000.
Siki, Z., Takács, B., 2021. Automatic recognition of ArUco codes in land surveying tasks. Baltic J. Modern Comput. 9, 115–125. https://doi.org/10.22364/BJMC.2021.9.1.06.
Siriani, A.L.R., Kodaira, V., Mehdizadeh, S.A., de Alencar Nääs, I., de Moura, D.J., Pereira, D.F., 2022. Detection and tracking of chickens in low-light images using YOLO network and Kalman filter. Neural Comput. Appl. 34, 21987–21997. https://doi.org/10.1007/s00521-022-07664-w.
The pandas development team, 2023. pandas-dev/pandas: Pandas (v2.0.3). Zenodo. https://doi.org/10.5281/zenodo.8092754.
Vagvolgyi, B.P., Jayakumar, R.P., Madhav, M.S., Knierim, J.J., Cowan, N.J., 2022. Wide-angle, monocular head tracking using passive markers. J. Neurosci. Methods 368, 109453. https://doi.org/10.1016/j.jneumeth.2021.109453.
Wang, K., Liu, K., Xin, H., Chai, L., Wang, Y., Fei, T., Oliveira, J., Pan, J., Ying, Y., 2019. An RFID-based automated individual perching monitoring system for group-housed poultry. Trans. ASABE 62, 695–704. https://doi.org/10.13031/trans.13105.
Wolf, S.W., Ruttenberg, D.M., Knapp, D.Y., Webb, A.E., Traniello, I.M., McKenzie-Smith, G.C., Leheny, S.A., Shaevitz, J.W., Kocher, S.D., 2023. NAPS: Integrating pose estimation and tag-based tracking. Methods Ecol. Evol. 2023, 2541–2548. https://doi.org/10.1111/2041-210X.14201.
Wubben, J., Fabra, F., Calafate, C.T., Krzeszowski, T., Marquez-Barja, J.M., Cano, J.C., Manzoni, P., 2019. Accurate landing of unmanned aerial vehicles using ground pattern recognition. Electronics (switzerland) 8, 1–16. https://doi.org/10.3390/electronics8121532.