Video-based Detection and Tracking with Improved Re-Identification Association for Pigs and Laying Hens in Farms

Qinghua Guo[1], Yue Sun[1], Lan Min[1], Arjen van Putten[2], Egbert Frank Knol[3], Bram Visser[4], T. Bas Rodenburg[2], J. Elizabeth Bolhuis[5], Piter Bijma[5] and Peter H. N. de With[1]
[1]*Department of Electrical Engineering, Eindhoven University of Technology, Eindhoven, The Netherlands*
[2]*Department of Animals in Science and Society, Utrecht University, Utrecht, The Netherlands*
[3]*Topigs Norsvin Research Center, Beuningen, The Netherlands*
[4]*Hendrix Genetics, Boxmeer, The Netherlands*
[5]*Department of Animal Sciences, Wageningen University & Research, Wageningen, The Netherlands*

Keywords: Animal Detection, Animal Tracking, Multi-Object Tracking Models.

Abstract: It is important to detect negative behavior of animals for breeding in order to improve their health and welfare. In this work, AI is employed to assist individual animal detection and tracking, which enables the future analysis of behavior for individual animals. The study involves animal groups of pigs and laying hens. First, two state-of-the-art deep learning-based Multi-Object Tracking (MOT) methods are investigated, namely Joint Detection and Embedding (JDE) and FairMOT. Both models detect and track individual animals automatically and continuously. Second, a weighted association algorithm is proposed, which is feasible for both MOT methods to optimize the object re-identification (re-ID), thereby improving the tracking performance. The proposed methods are evaluated on manually annotated datasets. The best tracking performance on pigs is obtained by FairMOT with the weighted association, resulting in an IDF1 of 90.3%, MOTA of 90.8%, MOTP of 83.7%, number of identity switches of 14, and an execution rate of 20.48 fps. For the laying hens, FairMOT with the weighted association also achieves the best tracking performance, with an IDF1 of 88.8%, MOTA of 86.8%, MOTP of 72.8%, number of identity switches of 2, and an execution rate of 21.01 fps. These results show a promising high accuracy and robustness for the individual animal tracking.

1 INTRODUCTION

While the demand for animal products increases, the people's attention for animal health and welfare is growing accordingly. Negative social interactions among animals, such as tail-biting in pigs and smothering in laying hens, threaten animal lives and reduce feeding efficiency, thereby increasing the ecological footprint of food production (Matthews, et al., 2016). Early detection of such negative behavior and interventions are essential, but challenging for farm staff due to changes over time and variances in environments, especially in large-scale farms (Matthews, et al., 2016), (Matthews, et al., 2017). The objective of the animal science community is to develop a one-health solution that jointly links human, animal, and environmental health (Kahn, 2017). To facilitate efficient breeding for animals with minimal occurrence of negative behavior, continuous monitoring of animals at a large scale is desirable for identifying damaging behavior. However, most animals are raised in groups, which causes inconvenience for observing individual animals. Therefore, automated and continuous individual tracking is needed.

In the past few years, several methods have been developed for automated animal monitoring. Radio frequency identification device (RFID) sensors have been widely used for tracking animals, which require the installation of RFID antennas at every location of interest in the housing environment, which can be deployed with tags, such as an ear tag for each pig (Kapun, et al., 2018), (Maselyne, 2016). However, sensors have risks of being destructed by the active behavior of animals. In addition, for large-scale commercial farms, RFIDs are expensive concerning the installation and retrieval of tags. Another rising field for animal tracking is based on videos, which are contactless, and can be more simply implemented (e.g. low-cost cameras) and maintained than RFID systems. Several studies have investigated the three-dimensional Kinect cameras monitoring from the top view with depth sensors (Mallick, et al., 2014), (Kim, et al., 2017). They are capable of monitoring animals

Guo, Q., Sun, Y., Min, L., van Putten, A., Knol, E., Visser, B., Rodenburg, T., Bolhuis, J., Bijma, P. and N. de With, P.
Video-based Detection and Tracking with Improved Re-Identification Association for Pigs and Laying Hens in Farms.
DOI: 10.5220/0010788100003124
In *Proceedings of the 17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2022) - Volume 4: VISAPP*, pages 69-78
ISBN: 978-989-758-555-5; ISSN: 2184-4321

through generated point clouds, while the range of depth sensors is too limited to address the entire area of a big pen (Matthews, et al., 2017). Additionally, the installation of top-view cameras could be difficult for large-scale farms. Therefore, the most common methods for monitoring animals are based on two-dimensional RGB cameras.

Recent advances in Artificial Intelligence (AI) provide radical new opportunities to monitor animal behavior through inexpensive and scalable strategies. The state-of-the-art Multi-Object Tracking (MOT) methods in deep learning include both two-stage and one-shot systems as shown in Figure 1. Two-stage

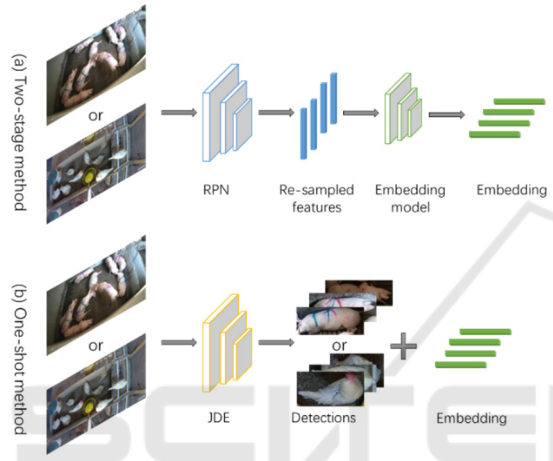

Figure 1: Layouts of both two-stage and one-shot tracking methods (Wang, et al., 2020).

methods firstly employ detectors, such as Faster R-CNN (Ren, et al., 2015) or YOLOv3 (Redmon et al., 2018) to localize objects in video frames, and then extract these features by an embedding model, such as Fast R-CNN (Girshick, 2015) for embedding learning. These two computations can adopt the most suitable model individually, achieving good performance on public pedestrian datasets (Wang, et al., 2020). However, separate detection and tracking tasks incur critical challenges on computation efficiency, while the execution time of embedding increases as the number of identities grows, which implies that two-stage methods are not optimal for real-time MOT in practice. To reduce computing time and enhance tracking efficiency, one-shot methods are proposed. Compared with two-stage tracking methods, one-shot methods combine object detection and embedding feature learning into a single deep network to reduce computation cost. In this way, detected objects and related appearance embeddings are learnt simultaneously in the network. The execution of the entire MOT procedure draws more attention than focusing on an association step only.

Our research aims at developing a 2D camera-based solution that leverages the state-of-the-art deep learning techniques for the automated detection and tracking of every individual pig/laying hen that is kept in large groups. In this work, we propose two one-shot video-based automated approaches for detecting and tracking individual pigs and laying hens. The first method is based on joint detection and embedding (JDE) network (Wang, et al., 2020), which is based on a one-shot concept of joint detection and tracking procedure. The second method FairMOT (Zhang, et al., 2020) is a network derived from the JDE by adding a re-ID embedding branch and addressing fairness issues to improve tracking performance. The proposed methods are evaluated using state-of-the-art metrics that provide multiple perspectives for assessing MOT (Heindl, 2017).

Our contributions to the improvement of the datasets are: (1) a pig dataset is manually annotated on 3,706 video frames including pigs with and without sprayed color marks, and (2) a laying hen dataset is created, containing 1,124 annotated frames of both white and brown laying hens. We apply two state-of-the-art methods to track individual objects for the two types of animals. Further contributions from the algorithmic side are as follows. An online association strategy is proposed based on animal characteristics, which efficiently reduces identity switches and enhances the tracking performance, for both JDE and FairMOT methods. Moreover, regarding the JDE, new clusters of anchor boxes are also learnt for each specific animal dataset, and the object numbers are constrained on each frame.

The sequel of this paper is as follows. Section 2.1 describes the data acquisition workflow, followed by Section 2.2 on the annotation method. Section 2.3 introduces the network architecture of the proposed methods. Section 2.4 describes the evaluation metrics, enabling visualization of the tracking performance. Section 3 illustrates experimental results, divided into Section 3.1 for pigs and Section 3.2 for laying hens. Section 4 discusses the findings of problems and the relevant future work accordingly. Section 5 discusses and concludes this paper.

2 METHODS

2.1 Dataset Description

2.1.1 Pig Dataset

All video recordings of pigs are collected at Volmer farm, Topigs Norsvin, Germany. Figure 2(a) shows a

sample frame illustrating the scene of the video recording. Pigs from in total eight pens are recorded, where each pen contains 10 or 11 pigs with or without sprayed color marks on the bodies. The group composition of the pigs usually remains unchanged unless situations of for example, sickness or injury occur. Most pens are set up with one single camera, while several remaining pens are equipped with double cameras. All the cameras film from the side views towards the pen's ground, covering the entire pen. The cameras used for recording are LOREX 4KSDAI168 with an image resolution of 1,280 × 720 pixels, and a frame rate of 15 fps. Pig videos are recorded continuously on a 24/7 basis, and each video is automatically generated and stored per hour.

2.1.2 Laying Hen Dataset

The videos of laying hens are collected at the farm of Utrecht University, the Netherlands. Figure 2(b) shows a sample frame illustrating the scene of the video recording. Laying hens from in total 24 pens are recorded, where each pen contains 8-10 laying hens. White and brown chickens are separated in different pens. The grouping of laying hens usually remains unchanged unless situations of e.g., sickness or injury occur. All pens are equipped with double cameras. The cameras film from the top and side views towards the pen's ground, coving the entire pen. The cameras used for recording are RLK8-800B4 with an image resolution of 2,304×1296 pixels, and a frame rate of 15 or 20 fps. Laying hens videos are recorded continuously during daytime.

2.2 Data Annotation

Video segments showing active animal movements are selected, followed by annotating the animal location in each video frame with consistent identity association for every animal. Computer Vision Annotation Tool (CVAT) (Intel, 2018) is used to label object location and also the situation of occluded objects. CVAT supports to save the frame ID, identity, bounding box location and size of the object.

2.3 Network Architecture Overview

2.3.1 Joint Detection and Embedding (JDE)

JDE network adopts DarkNet-53 (Redmon, et al., 2018) as the backbone network. It is based on Feature Pyramid Network (FPN) (Lin, et al., 2017), which provides possibilities to predict from multiple scales. As shown in Figure 3, at the beginning, the input video frames are transmitted forward through the

(a) A sample frame for pigs recorded at Volmer farm, Topigs Norsvin, Germany

(b) A sample frame for laying hens recorded at the farm of Utrecht University, the Netherlands

Figure 2: Sample frames for (a) pigs recorded at Volmer farm, Topigs Norsvin, Germany, and (b) laying lens recorded at the farm of Utrecht University, the Netherlands.

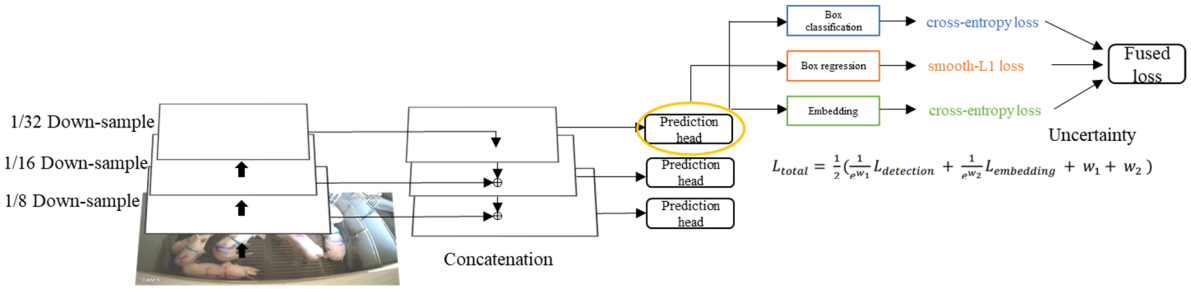

Figure 3: Explanation of JDE network architecture and prediction heads (Wang, et al., 2020).

backbone to obtain feature maps at three scales of down-sampling of 1/32, 1/16 and 1/8. The feature map, which supplies multi-task learning: box classification, box regression and embedding learning. The detection branch of JDE is based on the standard Region Proposal Network (RPN) (Ren, et al., 2015) map with the smallest size is up-sampled and fused with the feature map from the second smallest scale by a skip connection, similarly for other scales. Prediction heads are added to the fused feature maps at all scales. Each prediction head includes several convolutional layers and outputs a dense prediction with two modifications. First, we apply k-means clustering to the training dataset to recalculate 12 anchors, where each scale has 4 anchors. For the widely used pedestrian datasets in MOT, a filter condition is normally applied to constrain the object aspect ratio of 1:3 (width : height). We remove this constraint because more deformations are expected in the animal datasets. Second, we set the IoU threshold to distinguish foreground from background regions. In this way, false alarms can be suppressed especially for occluded cases. As depicted in Figure 3, the detection branch is covering two tasks: foreground/background classification with a cross-entropy loss, and bounding-box regression with a smooth L1 loss.

The learning procedure of appearance embedding in JDE is to derive a small distance measure for detected bounding boxes with the same identity, while bounding boxes with different identities have a large distance. JDE compares three types of loss functions to achieve this goal. The triplet loss (Schroff, et al., 2015) is feasible, but the training is unstable and its convergence is slow. A smooth upper bound of triplet loss (Sohn, 2016) is presented to alleviate issues caused by the triplet loss. It is similar to the cross-entropy loss, where all negative classes participate in the loss computation. However, the smooth upper bound of the triplet loss only considers sampled negative classes in the mini-batch. The experimental results with pedestrian datasets show that the cross-entropy loss gives the best results. Hence, the appearance embedding learning of JDE is based on using cross-entropy.

The way for combining detection loss and embedding loss is automatic loss balancing (Kendall, et al., 2018), based on the concept of task-independent uncertainty, which is calculated by

$$L_{total} = \frac{1}{2}\left(\frac{1}{e^{w_1}}L_{detection} + \frac{1}{e^{w_2}}L_{embedding} + w_1 + w_2\right) \quad (1)$$

where w_1 and w_2 are learnable parameters.

JDE adopts a simple and fast online association algorithm. Each tracklet consists of an appearance state and a motion state. The appearance affinity matrix is calculated by cosine similarity and the motion affinity matrix is computed using the Mahalanobis distance. A buffer pool is set for potential tracklets to the following association. For each frame, there are computations between all detections and tracklets in the buffer pool. The Hungarian algorithm (Kuhn, 1955) solves the linear assignment to output matched tracks, unmatched tracks and detections. A Kalman filter (Welch, et al., 1995) is used to update and predict the locations in the current frame from the existing tracklets.

Improved Re-identification Association: Figure 4 describes the workflow of an online association strategy. This procedure has three

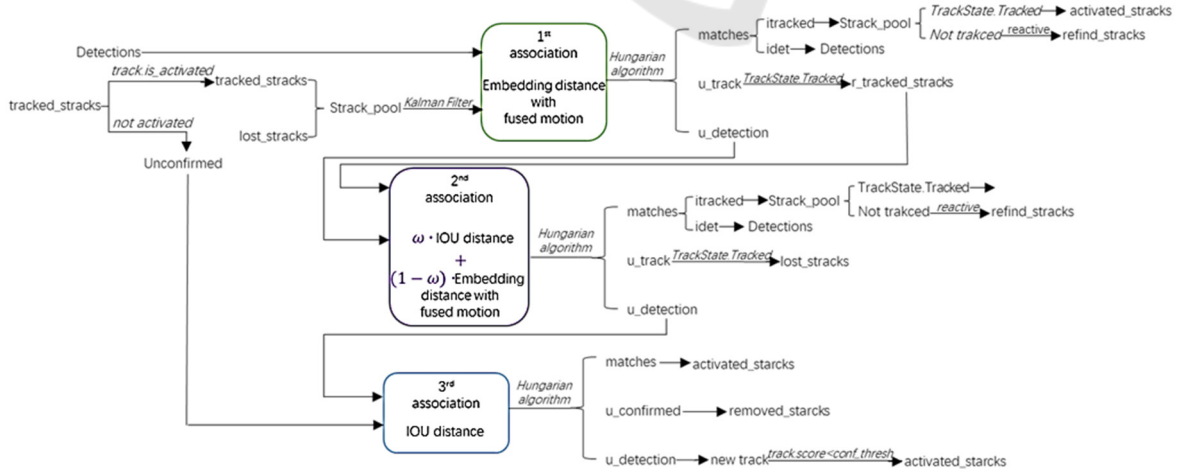

Figure 4: The online association strategy deployed both in JDE and FairMOT. We improve the second association with the weighted IoU distance and embedding distance with fused motion, instead of only considering the IoU distance at the second association in JDE and FairMOT.

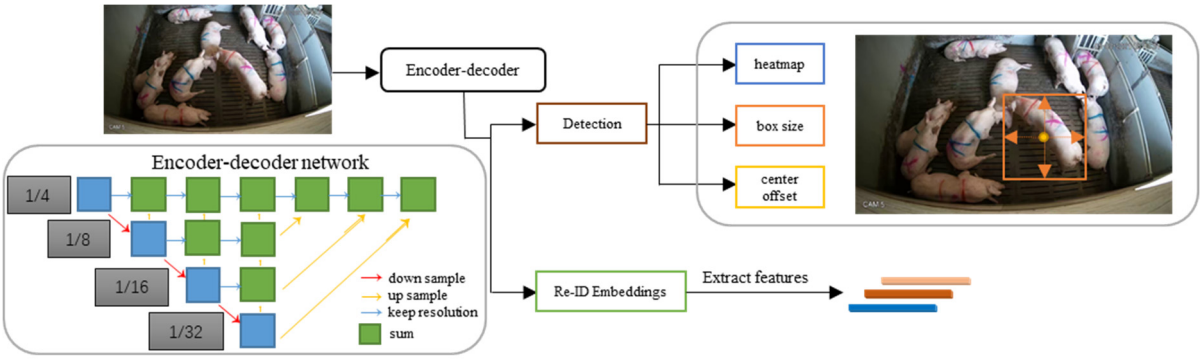

Figure 5: Explanation of the FairMOT (Zhang, et al., 2020) network architecture and prediction heads.

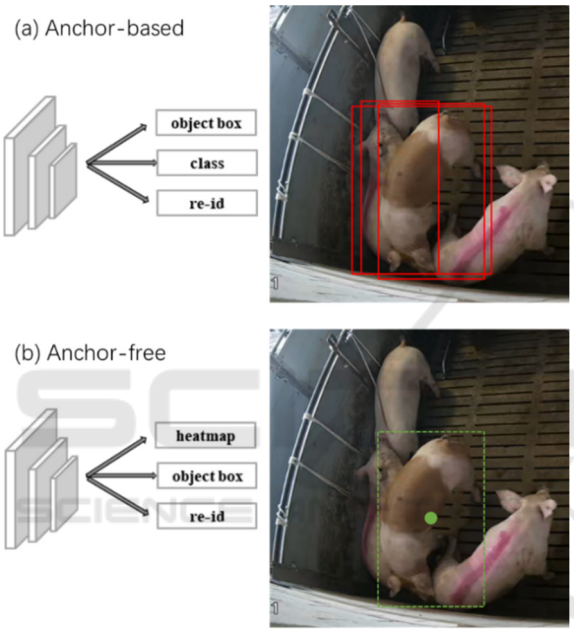

Figure 6: Comparison between (a) anchor-based (JDE) and (b) anchor-free method (FairMOT) (Zhang, et al., 2020).

association steps in total. The first association is related to the embedding distance with fused motion.

After calculating by the Hungarian algorithm (Kuhn, 1955), the unmatched tracks and detections are further imported to the second association. We introduce an improved strategy, which considers the weighted distance between the IoU and fused embedding with motion, instead of only relying on the IoU distance after the first embedding comparison. Because animal behavior is faster and more deformable than pedestrians, the comparison of appearance embeddings is more reliable. In the third ID association step, the IoU distance is adopted to handle the unconfirmed tracks, which are usually tracks with only one initial frame. A buffer pool is used for storing lost tracks, and the tracks are removed when they have been lost for more than a certain frame count (threshold). Finally, the outputs combine all followed tracks, activated tracks and refined tracks.

In addition, we present a way to limit object numbers in each frame by Non-Maximum Suppression (NMS) (Neubeck, et al., 2006). According to the amount of objects for tracking, we keep the same amount of NMS indices.

2.3.2 FairMOT

The backbone network used in FairMOT is ResNet-34, which trades-off tracking performance and computing time. To fuse multi-layer features like JDE, a developed version of Deep Layer Aggregation (DLA) (Zhou, et al., 2019) is attached to the backbone as shown in Figure 5. The development adds more skip connections between multiple scales, which is similar to the FPN. Moreover, there are deformable convolution layers in all up-sampling stages, which enables dynamic adjustment among object scales and poses. The entire network is called DLA-34.

Compared with JDE, FairMOT addresses three unfair issues caused by anchors, features, and feature dimensions. Figure 6 illustrates the unfairness caused by the anchor-based method in JDE. As shown in Figure 6(a), all active anchors around the object center are considered as candidates of Re-ID features. These adjacent anchors have high possibilities to be confirmed as the same identities if their IoU value is large enough, which results in suboptimal extracted features. For instance, Figure 6(a) shows that three anchors are predicted as the same identity. FairMOT solves this unfairness by extracting the Re-ID feature only from the center of the object (see Figure 6(b)). In addition, FairMOT improves the setting of the feature dimension, while the performance is higher when the network learns lower-dimensional features.

The detection module in FairMOT is based on the CenterNet, but also combined with anchor-free methods. It leaves out the steps for computing clusters from all bounding boxes. As can be observed in Figure 5, three parallel heads contribute to the detection branch. The heatmap head predicts the locations of the object centers with a focal loss. The box-offset head and the box-size head are responsible for more accurate localization and estimating the height and width of the target box, optimized by the L1 loss.

As shown in Figure , FairMOT introduces a re-ID branch to generate object features, aiming at distinguishing different objects. The re-ID features are extracted from the feature map, which are derived from a convolution layer with 128 kernels based on the backbone network.

The automated loss balancing and online association strategy in FairMOT are the same as used in the JDE network. We explore a weighted strategy in the FairMOT approach, which is expected to reduce the identity switches during tracking.

2.4 Evaluation Metrics

The proposed methods are evaluated using the metrics derived from the MOT challenge based on a pedestrian dataset (Dendorfer, et al., 2019), combined with evaluation metrics used in JDE (Wang, et al., 2020) and FairMOT methods (Zhang, et al., 2020), (Bernardin, et al., 2008). These metrics are employed and listed in Table 1 and supplemented with the IDF1 metric (Ristani, et al., 2016) to evaluate the overall tracking performance. Table 1 illustrates all terms for evaluating MOT systems. The upward arrow means a higher value of this term is desired and the downward arrow indicates that a result of a lower value is better.

Table 1: Evaluation metrics for the proposed methods.

Metric	Description
MOTA↑	Multi-Object Tracking Accuracy. This measure combines three error sources: false positives, missed targets, and identity switches.
MOTP↑	Multi-Object Tracking Precision. The misalignment between the annotated and the predicted bounding boxes.
MT↑, PT, ML↓	Number of mostly tracked, partially tracked, and mostly lost trajectories.
IDF1↑	ID F1 score. The ratio of correctly identified detections over the average number of ground-truth and computed detections.
IDs↓	Number of identity switches.
FPS↑	Runtime, frame per second.

Table 2: Summary of training data information for pigs.

Rec. date	Sprayed marks	No. of pens	No. of identities	No. of frames	No. of bounding boxes
20200820	no	8	87	1,737	36,019
20201205	yes	6	66	600	
20210105	yes	6	65	561	
20210205	yes	3	33	303	
20210305	yes	1	10	101	
Overall	with:without = 10:11	24	261	3,302	

Table 3: Summary of testing data information for pigs.

Rec. No.	Rec. date	Sprayed marks	No. of identities	Duration (min : s)	No. of frames	Bounding boxes	Conditions
R1_pig	20210205	yes	11	3:20	101	1,111	Limited active movements; Pigs are not too close to each other
R2_pig	20210415	yes	11	3:20	101	1,111	Partial active movements; one pig is mostly occluded in a few seconds
R3_pig	20210420	no	11	3:20	101	1,111	Partial active movements; some pairs of pigs are very close
R4_pig	20210420	no	11	3:20	101	1,111	Partial active movements; several pigs are stacked together

Table 4: Comparison between the original association method and the proposed weighted strategy on the pig dataset.

Testing	Method	Association	IDF1↑	MOTA↑	IDs↓
All testing data	JDE	Original	82.0	**90.1**	54
		Weighted	**82.9**	89.9	**36**
	Fair-MOT	Original	89.7	90.8	18
		Weighted	**90.3**	90.8	**14**

Table 5: Comparison of tracking results from JDE and FairMOT on the pig testing set.

Method	Test recording	IDF1↑	MOTA↑	MOTP↑	GT	MT↑	PT	ML↓	IDs↓	FPS↑
JDE	R1_pig	82.8	93.5	85.5	11	11	0	0	7	15.98
	R2_pig	88.9	**92.0**	79.1	11	**11**	0	0	5	15.83
	R3_pig	81.5	**89.7**	78.4	11	**11**	0	0	6	15.93
	R4_pig	78.4	84.4	78.8	11	9	2	0	18	15.70
	Overall	82.9	89.9	80.5	44	42	2	0	36	15.86
FairMOT	R1_pig	**91.2**	**94.4**	**87.8**	11	11	0	0	**2**	**20.57**
	R2_pig	**94.3**	90.5	**82.2**	11	10	1	0	**1**	**20.41**
	R3_pig	**88.1**	87.3	**81.8**	11	10	1	0	**4**	**20.50**
	R4_pig	**87.8**	**91.1**	**82.7**	11	**11**	0	0	**7**	**20.42**
	Overall	**90.3**	**90.8**	**83.7**	44	42	2	0	**14**	**20.48**

3 EXPERIMENTAL RESULTS

3.1 Pig Experiments

3.1.1 Dataset Description

Our manually annotated datasets of pigs are divided into two parts, which are the training dataset described in Table 2, and the testing dataset from Table 3. Pig videos are recorded at a frame rate of 15 fps. A frame step of 30 frames (2 seconds) is taken during annotation to output one frame. All images are selected from daytime in an uncontrolled farming environment. The training dataset shown in Table 2 consists of 3,302 frames including 36,019 annotated bounding boxes from 24 pens in 5 days, of which the recording dates are separated from each other for at least one month. The frame ratio between pigs with/without sprayed color marks is about 10/11. Four videos with the same length and the same number of pigs are used for testing as depicted in Table 3. For evaluating the generalizability of our models, videos in various conditions are selected according to the activity levels of pig movements, occlusion or occurrence of group stacking.

3.1.2 Implementation Details

The backbone network of JDE is DarkNet-53 (Redmon, et al., 2018). Twelve clusters of anchor boxes are derived from all training bounding boxes by a k-means clustering method. Three key parameters are determined - learning rate, batch size and epoch by smartly choosing the training parameters based on controlled experiments, to yield the best convergence and the highest accuracy. The training model is based on a learning rate of 0.001, optimizing under standard SGD. The training is performed for 30 epochs with a batch size of unity. The input video frames are resized to $1{,}088 \times 608$ pixels.

The backbone network of FairMOT is DLA-34 (Zhang, et al., 2020). The initialized weights are pre-trained on the pedestrian dataset (Zhang, et al., 2020) by the DLA-34 network. The training model starts with a learning rate of 0.0001, optimized with the Adam optimizer. The training is performed for 50 epochs with a batch size of 2. The input video frames are also resized to $1{,}088 \times 608$ pixels.

All the experiments are carried out on a GeForce GTX 1080 GPU and an Xeon E5-1650 v4 CPU.

3.1.3 Results on Improving re-ID

We assess the models on all testing data with JDE and FairMOT methods. Table 4 shows a comparison between the original association method and the proposed weighted association strategy, where the results of IDF1, MOTA and ID switches are shown. The value of MOTA decreases by 0.2 percent by the weighted association. However, the weighted strategy outperforms the original association, as

shown by the other two metrics. The overall results of FairMOT are better than JDE, especially with respect to identity switches.

Table 4 demonstrates that the weighted association is effective. Hence, the following evaluations are all on the models with the weighted association step. Evaluation metrics on separate test recordings are then calculated (see Table 5). Most results show that FairMOT outperforms JDE, especially in terms of the execution time (FairMOT executes about 5 fps faster than JDE). The average values of IDF1 and MOTA are increased relatively by 7.4% and 0.7%. Another essential term in MOT refers to the identity switches, where the obtained result of FairMOT is lower than half of that achieved by JDE.

3.2 Laying Hen Experiments

3.2.1 Dataset Description

The laying hen dataset includes the training dataset (see Table 6 for details) and the testing dataset (see Table 7 for details). Laying hen videos are recorded at a frame rate of 15 or 20 fps. A frame step of 15 or 20 frames (1 second) is chosen to output one annotated frame. All images are selected from daytime with uncontrolled environmental conditions. The training dataset as shown in Table 6 consists of 2,563 frames including 21,708 annotated bounding boxes from 8 pens in 4 days. The frame ratio between white/brown laying hens is around 7/10. Two videos with the same length and the same number of laying hens are used for testing, as shown in Table 7. For evaluating the generalizability of the proposed models, videos in different conditions are selected according to feather color, activity levels of laying hen movements and occlusion occurrences.

3.2.2 Implementation Details

The experimental settings of laying hen training on JDE are similar to the training pig data, except for the learning rate, which is set to 0.0001.

The training procedure for the laying hen dataset on FairMOT is also similar to training pig data, but performed for 100 epochs.

3.2.3 Results on Improving re-ID

The proposed models are evaluated on the testing data. Table 8 provides a comparison between the original association method and the proposed weighted association strategy, where the results illustrate the values of IDF1, MOTA and ID switches. The results demonstrate that the weighted strategy is feasible to improve the tracking results for both models.

The performances of two tracking methods are improved by applying the weighted association, as shown in Table 8. Hence, the following evaluations are based on the models with the weighted association

Table 6: Summary of training data property for the laying hens.

Rec. date	Color	No. of pens	No. of identities	No. of frames	No. of bounding boxes
20210308	brown	1	10	288	21,708
20210318	white, brown	2	88	1,085	
20210321	white, brown	2	19	196	
20210720	brown	3	35	994	
Overall	white:brown= 7 : 10	8	152	2,563	

Table 7: Summary of testing data property for laying hens.

Rec. No.	Rec. date	Color	No. of identities	Duration (min : s)	No. of frames	Bounding boxes	Conditions
R1_hen	20210318	white	9	2:30	151	1,359	Mostly active movements; hens fly several times.
R2_hen	20210720	brown	9	2:30	151	1,359	Partial active movements; One hen is completely occluded.

Table 8: Comparison between original association method and the proposed weighted strategy on the laying hen dataset.

Testing	Method	Association	IDF1↑	MOTA↑	IDs↓
All testing data	JDE	Original	**83.7**	84.7	5
		Weighted	83.1	**85.1**	5
	FairMOT	Original	87.1	86.5	4
		Weighted	**88.8**	**86.8**	**2**

Table 9: Comparison of tracking results from JDE and FairMOT on laying hen testing set.

Method	Test recording	IDF1↑	MOTA↑	MOTP↑	GT	MT↑	PT	ML↓	IDs↓	FPS↑
JDE	R1_hen	84.0	85.5	77.6	9	9	0	0	2	18.16
	R2_hen	82.2	84.8	**83.7**	9	8	0	1	3	18.16
	Overall	83.1	85.1	**80.6**	18	17	0	1	5	18.16
FairMOT	R1_hen	**85.1**	**87.2**	**78.5**	9	9	0	0	2	**21.01**
	R2_hen	**92.8**	**86.4**	77.5	9	8	0	1	**0**	**21.01**
	Overall	**88.8**	**86.8**	72.8	18	17	0	1	**2**	**21.01**

algorithm. The evaluation metrics on separate test recordings are obtained (see Table 9), where for most metrics, FairMOT outperforms JDE, especially in terms of the execution time (FairMOT is about 3 fps faster than JDE). The average values of IDF1 and MOTA are relatively increased by 5.7% and 1.7%. Again, the number of identity switches from FairMOT is less than half of that with JDE. Especially, FairMOT achieves zero identity switches for the second recording. It can be observed that the execution time of FairMOT is lower than for JDE, which is beneficial for real-time multi-object tracking. If more data is added, we still expect similar results, but with more reliability in the comparison.

4 DISCUSSION AND CONCLUSION

In this paper, we investigate two state-of-the-art automated multi-object tracking methods on animal datasets. Manual annotation of two types of animals are collected: 3,706 frames of pigs with / without sprayed body marks, and 2,865 frames of white / brown laying hens. The models are evaluated on 4 pig videos, each lasting 3 minutes and 20 seconds, and 2 laying hen videos, each lasting 2 minutes and 30 seconds. Each recording has different challenging conditions such as occlusion, active and high-speed movements. In this way, the generalization and robustness of the tracking models are evaluated. The execution time on JDE is 15~18 fps, while FairMOT can achieve more than 20 fps. We have proposed a weighted association strategy to improve the association algorithm of animal re-ID, which increases the performance of IDF1 by 1.7% at most, MOTA by 0.4% at most and reduces the identity switches by 18 at most.

Overall, the evaluation metrics of JDE on the pig dataset result in an IDF1 of 82.9%, MOTA of 89.9%, MOTP of 80.5%, number of identity switches of 36 and a rate of 15.86 fps. FairMOT deployed on the pig dataset results in an IDF1 of 90.3%, MOTA of 90.8%, MOTP of 83.7%, number of identity switches of 14 and an execution speed of 20.48 fps.

For the laying hen dataset, JDE leads to an IDF1 of 83.1%, MOTA of 85.1%, MOTP of 80.6%, number of identity switches of 5 and a execution speed of 18.16 fps. FairMOT yields an IDF1 of 88.8%, MOTA of 86.8%, MOTP of 72.8%, number of identity switches of 2 and an execution speed of 21.01 fps.

Considering the manual annotation effort, the procedure for collecting appropriate annotation is rather slow. However, better performance is expected when training on more available data is possible. Considering the difference in moving speed of animals, we have adopted an annotation step of 2 seconds for pigs and 1 second for laying hens to improve annotation efficiency. Continuous annotation is expected to yield a more precise tracking system. Additionally, our ultimate goal is to achieve good real-time animal tracking, so longer video recordings in different conditions are required to be annotated for both model development and more thorough evaluation. After achieving sufficient tracking performance, we will also optimize and trade-off the execution time of the system.

The adopted backbone network in JDE is DarkNet-53 for object detection, which is based on the third version of YOLO. In recent years, YOLO has already been developed and implemented into Version 5. Future work will involve to incorporate the latest YOLOv5 into the JDE model in order to verify its efficiency. Similar work should also be performed for the FairMOT architecture.

ACKNOWLEDGEMENTS

This publication is part of the project IMAGEN [P18-19 project 1] of the research programme Perspectief, which is financed by the Dutch Research Council (NWO).

REFERENCES

Bernardin, K., & Stiefelhagen , R. (2008). Evaluating Multiple Object Tracking Performance: The CLEAR MOT Metrics. *EURASIP Journal on Image and Video Processing, 2008*, 1-10.

Chung, M. K., Lee L. Eckhardt, & Lin Y. Chen. (2020). Lifestyle and Risk Factor Modification for Reduction of Atrial Fibrillation: A Scientific Statement From the American Heart Association. *Circulation*. doi:10.1161/ CIR.0000000000000748

Costa, Madalena, Goldberger, Ary L, Peng, & C.-K. (2002, 7). Multiscale Entropy Analysis of Complex Physiologic Time Series. Phys. Rev. Lett., 89(6), 068102. Retrieved from https://link.aps.org/doi/ 10.1103/PhysRevLett.89.068102

Dendorfer, P., Osep, A., & Leal-Taixé, L. (n.d.). *CVPR 2019 Tracking Challenge Results*. Retrieved from https://motchallenge.net/

FG, C., Aliot E, & Botto GL. (2008). Delayed rhythm control of atrial fibrillation may be a cause of failure to prevent recurrences: reasons for change to active antiarrhythmic treatment at the time of the first detected episode. *Europace*. doi:10.1093/europace/eum276

Girshick, R. (2015). Fast R-CNN. *ICCV 2015.*

Goldberger, A., Amaral, L., Glass, L., & Hausdorff, J. (2017, 2 1). *AF Classification from a Short Single Lead ECG Recording - The PhysioNet Computing in Cardiology Challenge 2017*. (PhysioBank, PhysioToolkit, and PhysioNet) Retrieved from https://physionet.org/ content/challenge-2017/1.0.0/

Griffin, D. W., & Jae S. Lim. (1984). Signal Estimation from Modified Short-Time Fourier Transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 236-243.

Heindl, C. (2017). *Benchmark multiple object trackers (MOT) in Python*. Retrieved from https://github.com/ cheind/py-motmetrics

Intel. (2018). *Powerful and efficient Computer Vision Annotation Tool (CVAT)*.

Kahn, L. H. (2017). Perspective: The one-health way. *Nature*, 543, S47.

Kapun, A., Felix, A., & Eva, G. (2018). Activity analysis to detect lameness in pigs with a UHF-RFID system. *10th International Livestock Environment Symposium*.

Kendall, A., Gal, Y., & Cipolla, R. (2018). Multi-Task Learning Using Uncertainty to Weigh Losses for Scene Geometry and Semantics. *CVPR.*

Kim, J., Chung, Y., Choi, Y., Sa, J., Kim, H., Chung, Y., . . . Kim, H. (2017). Depth-Based Detection of Standing-Pigs in Moving Noise Environments. *Sensors, 17*(12), 2757.

Kuhn, H. W. (1955). The hungarian method for the assignment problem. In *Naval Research Logistics Quarterly* (pp. 83-97).

Lee, G. R., Ralf Gommers, Filip Wasilewski, Kai Wohlfahrt, & Aaron O'Leary. (2019). PyWavelets: A Python package for wavelet analysis. *Journal of Open Source Software*, 4(36), 1237. doi:https://doi.org/ 10.21105/ joss.01237

Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. (2017). Feature Pyramid Networks for Object Detection. *CVPR 2017.*

Lip, G., L. Fauchier, & S.B. Freedman. (2016). Atrial fibrillation. *Nat Rev Dis Primers 2*. doi:https: //doi.org/10.1038/nrdp.2016.16

Mallick, T., Das, P. P., & Majumdar, A. K. (2014). Characterizations of Noise in Kinect Depth Images: A Review. *IEEE Sensors Journal, 14*(6), 1731-1740.

Maselyne, J. (2016). Measuring the drinking behaviour of individual pigs housed in group using radio frequency identification (RFID). *Animal*, 1557-1556.

Matthews, S. G., Miller, A., James, C., Llias, K., & Thomas, P. (2016). Early detection of health and welfare compromises through automated detection of behavioural changes in pigs. *The Veterinary Journal, 217*, 43-51.

Matthews, S., A.L., M., & Thomas, P. (2017). Automated tracking to measure behavioural changes in pigs for health and welfare monitoring. *Scientific Reports*, 7.

Munger, T. M., Li-Qun Wu, & Win K. Shen. (2014). Atrial fibrillation. *Journal of Biomedical Research*. doi:10.7555/jbr.28.20130191

Neubeck, A., & Gool, L. V. (2006). Efficient non-maximum suppression. *IEEE.*

Page, R. L., W E Wilkinson, W K Clair, E A McCarthy, & E L Pritchett. (1994). Asymptomatic arrhythmias in patients with symptomatic paroxysmal atrial fibrillation and paroxysmal supraventricular tachycardia. *Circulation*. doi:https://doi.org/10.1161/01.CIR.89.1.224

Redmon, J., & Farhadi, A. (2018). *Yolov3: An incremental improvement.*

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *Advances in Neural Information Processing Systems* (pp. 91-99).

Ristani, E., Solera, F., Zou, R., Cucchiara, R., & Tomasi, C. (2016). Performance Measures and a Data Set for Multi-Target, Multi-Camera Tracking. *ECCV 2016.*

Schroff, F., Kalenichenko, D., & Philbin, J. (2015). FaceNet: A Unified Embedding for Face Recognition and Clustering. *CVPR.*

Sejdić, E., Igor Djurović, & Jin Jiang. (2009). Time-frequency feature representation using energy concentration: An overview of recent advances. *Digital Signal Processing*, 19 (1): 153–183. doi:https://doi.org/10.1016/j.dsp.2007.12.004

Sohn, K. (2016). Improved Deep Metric Learning with Multi-class N-pair Loss Objective. *NIPS.*

T.Inouye. (1991). Quantification of EEG irregularity by use of the entropy of the power spectrum. *Electroencephalography and clinical neurophysiology, 79*(3), 204-210.

Wang, Z., Zheng, L., Liu, Y., Li, Y., & Wang, S. (2020). Towards Real-Time Multi-Object Tracking. *2020 European Conference on Computer Vision.*

Welch, G., & Bishop, G. (1995). *An introduction to the kalman filter.*

Zhang, Y., Wang, C., Xinggang, W., Wenjun, Z., & Wenyu, L. (2020). FairMOT: On the Fairness of Detection and Re-Identification in Multiple Object Tracking. *2020 Conference on Computer Vision and Pattern Recognition.*

Zhou, X., Wang, D., & Krähenbühl, P. (2019). *Objects as Points.* arXiv.