Honeybee Re-identification in Video: New Datasets and Impact of Self-supervision

Jeffrey Chan[1], Hector Carrión[2], Rémi Mégret[2], José L. Agosto Rivera[3] and Tugrul Giray[3]

[1]*Department of Mathematics, University of Puerto Rico, Río Piedras Campus, Puerto Rico*

[2]*Department of Computer Science, University of Puerto Rico, Río Piedras Campus, Puerto Rico*

[3]*Department of Biology, University of Puerto Rico, Río Piedras Campus, Puerto Rico*

Keywords: Re-identification, Contrastive Learning, Self-supervised Learning, Animal Monitoring.

Abstract: This paper presents an experimental study of long-term re-identification of honeybees from the appearance of their abdomen in videos. The first contribution is composed of two image datasets of single honeybees extracted from 12 days of video and annotated with information about their identity on long-term and short-term scales. The long-term dataset contains 8,962 images associated to 181 known identities and used to evaluate the long-term re-identification of individuals. The short-term dataset contains 109,654 images associated to 4,949 short-term tracks that provide multiple views of an individual suitable for self-supervised training. A deep convolutional network was trained to map an image of the honeybee's abdomen to a 128 dimensional feature vector using several approaches. Re-identification was evaluated in test setups that capture different levels of difficulty: from the same hour to a different day. The results show using the short-term self-supervised information for training performed better than the supervised long-term dataset, with best performance achieved by using both. Ablation studies show the impact of the quantity of data used in training as well as the impact of augmentation, which will guide the design of future systems for individual identification.

1 INTRODUCTION

The United Nations estimated that around 1 million animals and plants are threatened with extinction causing a dangerous decline of species (UN Press material, 2019). Active monitoring of endangered species can prevent extinction by the early detection of threats and studies of survival behaviors. Current monitoring systems are categorized as intrusive (Boenisch et al., 2018; Mégret et al., 2019) and non-intrusive (Bozek et al., 2021; Romero-Ferrero et al., 2019). Intrusive monitoring involves attaching a marker to individuals to reduce monitoring to the detection and identification of markers. This approach simplifies the analysis, but is restricted to controlled environments with access to the individuals in advance to perform marking. It present the advantage of providing individualized analysis of behavior patterns, which provide much finer grained information for detailed assessment of animal health, social behavior and division of labour amongst others. On the other hand, non-intrusive monitoring consists of placing a camera trap without any marker. In this case, the detection and tracking can then be performed using computer vision algorithms, leaving identification to a set of experts if the number of events detected is small enough. Manual identification is an arduous and time-consuming task, which then require automation for large time spans or if many individuals are to be monitored, which is the case for honeybees.

Recently, (Romero-Ferrero et al., 2019) developed a method for markerless tracking of groups of animals in laboratory conditions where the animal always stays in the camera field of view. They trained a re-identification model using tracking information to incrementally build appearance models of each individual to solve ambiguities during crossings. An incremental approach to build individual appearance models from initial partial trajectories was used in (Bozek et al., 2021) to solve track interruptions from images of honeybees inside an observation colony, this time relaxing the constraint of fixed number of individuals. Re-identifying animals in their natural habitat is particularly challenging because an individual can decide to go out of the field of view for an indefinite amount of time. For this reason, it requires robust models to connect tracks with a significant time gap in between and undefined number of individuals.

Chan, J., Carrión, H., Mégret, R., Rivera, J. and Giray, T.
Honeybee Re-identification in Video: New Datasets and Impact of Self-supervision.
DOI: 10.5220/0010843100003124
In *Proceedings of the 17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2022) - Volume 5: VISAPP*, pages 517-525
ISBN: 978-989-758-555-5; ISSN: 2184-4321

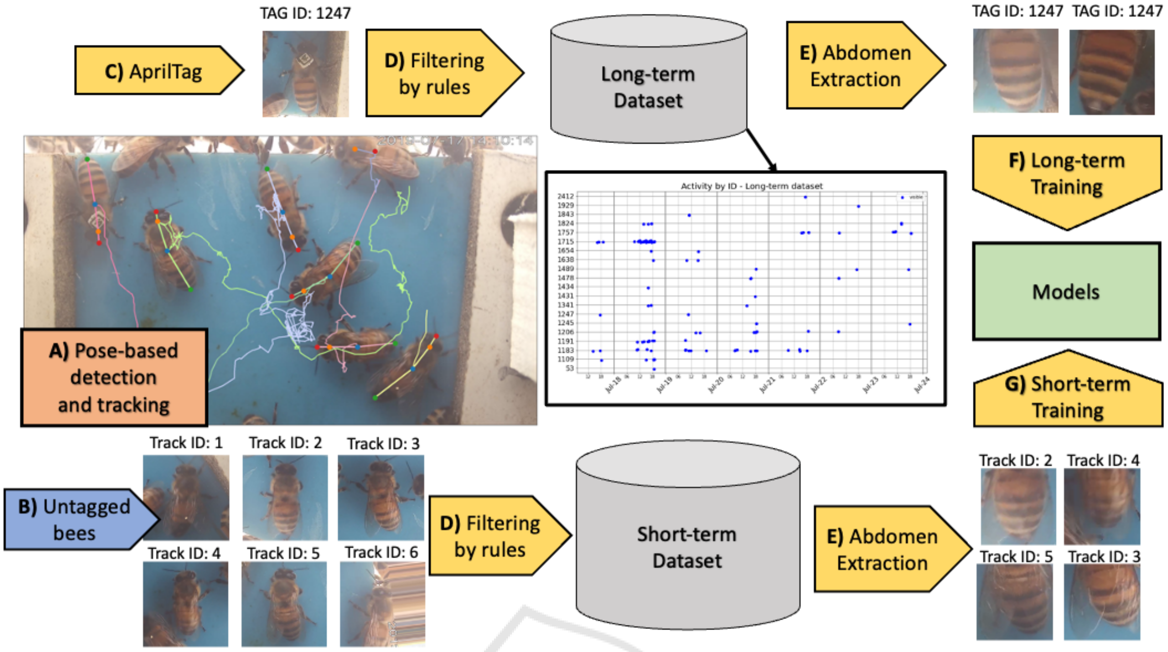

Figure 1: General architecture of the creation of the short-term and long-term re-identification datasets. A) Pose-based detector detects head, neck, waist, and abdomen tip and performs short-term tracking based on the waist. B) Angle compensated body extraction of untagged bees. C) Decode tags using AprilTag and body extraction with angle compensation of tagged bees. D) Filter detections using rules based on abdomen size, abdomen angle, and closest bee. E) Abdomen extraction, remove the upper body to avoid data leaks from tagged images. G) Long-term training dataset where the identity class is the tag id. H) Short-term training dataset where the identity class is the track id.

Using contrastive learning, (Schneider et al., 2020) achieved re-identification performance beyond humans capabilities on datasets of tigers, fruit fly, monkeys, and whales. One of the key to success in this case is the large number of images annotated with their groundtruth identity, which where carefully obtained from experts. Unfortunately, such manual labeling of identities is much more challenging for some species, such as honeybees, due to the large number of individuals and the lack of human experts that can perform this task.

Capturing the variations of appearance of the same individual can also be obtained using detection and tracking algorithms, but this is limited to monitor animals while the individual stays in the camera's field of view. This can capture short-term variations by associating contiguous instances to the same identity and learn invariance to pose changes, rotation, and deformation.

Because of the difficulty of collecting large-scale data with supervised identity annotation, we investigate in this paper how to leverage the short-term tracking information as self-supervised training information and evaluate its impact on the performance in long-term re-identification of honeybees.

The paper is organized as follows. In section 2, we review the related work in terms of methodology and application of re-identification to animals. In section 3, we present the design and building of two honeybee re-identification datasets that will be shared with to the community and that will be used for a detailed evaluation of performances. In section 4, we introduce the models, training and evaluation procedures. In section 5, we show the experimental results and discuss their implications for the development of improved re-identification approaches, before concluding in section 6.

2 RELATED WORK

2.1 Animal Re-Identification

Human Re-identification is a well-known task in the field of computer vision. The community had been very active for years thanks to the availability of massive labeled datasets such as Market-1501 (Zheng et al., 2015), and CUHK03 (Li et al., 2014) that enables the development of specialized methods for human re-identification. Unfortunately, the animal re-identification community had not been able to reach the same performance. A major factor is the availability of identity annotated datasets. Animal datasets

Table 1: Organization and statistics of the contributed datasets.

Dataset	Split	# individuals	# images	# tracks	Mean images per tracks	Mean tracks per id
short-term	train	-	109,654	4,949	22.15	-
long-term	train	181	3,777	801	4.71	4.42
	valid	66	1,909	309	6.17	4.68
	test	126	3,276	696	4.70	5.52

have their unique characteristics, animals do not wear clothes or makeup, but individuals may look very similar to each other, making the annotation an arduous and time-consuming task that even sometimes is unfeasible for experts.

Although animal datasets are hard to collect, recent efforts have led to the collection of medium-size datasets for species such as tigers (Li et al., 2019), elephants (Körschens et al., 2018), cattle (Gao et al., 2021; Bergamini et al., 2018), and primates (Deb et al., 2018; Brust et al., 2017; Schofield et al., 2019). Most of these datasets have focused on supervised learning by involving experts to label the identity of the individuals. For species with a large number of individuals such as honeybees, data collection can capture hundreds or even thousands of identities quickly, but annotation cannot be performed by experts. Therefore, this work explores self-supervision toward a re-identification of honeybees.

Recently the Cows2021 dataset (Gao et al., 2021) used tracking and self-supervised learning to help to annotate more individuals for the dataset based on their color patterns. The authors used triplet loss, sampling the positive pair from the same track, and the negative example from a different video to train an invariant feature space from 301 short videos, reaching 0.57 ID accuracy on a different test set with 182 individuals.

2.2 Self-supervised Learning

The success of a deep model on a visual task depends on learning suitable features for the downstream task, such as image classification, object detection, or re-identification. Pretraining had become a crucial component of the model training to achieve state-of-the-art performance. In pretraining, the model is optimized to perform a similar task to learn initial relevant features before fine-tuning with the downstream task. This similar task may benefit from a massive annotated dataset to train the network, such as the ImageNet dataset. For fine-grained tasks where a massive dataset is not available self-supervised learning is used. Self-supervised learning aims to pre-train a network with a pretext task that does not require manually annotated labels (Misra and Maaten, 2020; Noroozi and Favaro, 2016; Chen et al., 2020a). SimCLRv2 had been shown to outperforms standard supervised training, even when fine-tuning with only 10% of the labels (Chen et al., 2020b). These results motivate the approach we propose, where we combine data augmentation and tracking as a generator for pseudo labels to learn visual features relevant to honeybees' re-identification.

3 DESIGN OF THE DATASET

The purpose of this dataset is to evaluate the re-identification of unmarked bees on a long-term setup using two training modalities: 1) short-term training dataset which captures an individual in a short period; and 2) long-term training dataset which has annotations of individuals on long-term period. Figure 1 shows an overview of the pipeline for the extraction of both short-term and long-term datasets.

3.1 Extraction of Individual Images

The raw video data was collected using a camera at the entrance of a colony recording honeybees' activities over multiple weeks. The videos were recorded at 20 fps, with quality of 1440x2560 pixels from July 17-24 and August 1-4 from 8 am to 6 pm. A subset of the honeybees was tagged several days prior to recording to ensure they would perform foraging trips and be visible in the monitored entrance. A paper tag was attached to their body, containing unique AprilTag barcodes (Wang and Olson, 2016). The data collection of tagged bees is scarce and requires delicate manipulation of the honeybees. Meanwhile, the data collection of untagged bees is automatic and massive as it only depends on individuals to appear in the camera field of view.

For all videos, the bee pose estimator (Rodriguez et al., 2018) was used to detect the skeleton. We used a modified skeleton template that includes the head, neck, waist, and abdomen tip. The annotations to train the pose estimation model were made using the SLEAP annotation interface (Pereira et al., 2020) to annotate 98 frames from one-hour video manually. The waist keypoint is used as a reference point for tracking, which is performed with a Hungarian al-

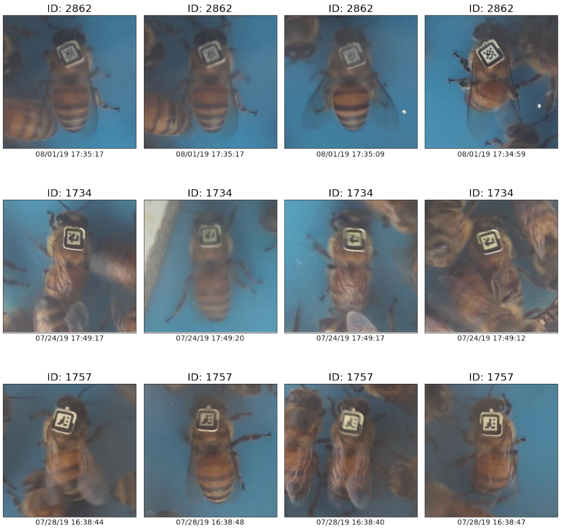

Figure 2: Example images of the raw dataset. Some images exhibit abdomen curling, crowded images and occlusions that were filtered out based on pose detection data.

gorithm. The neck and waist key points are used to compute the angle to normalize the body in the image extraction. Figure 1A shows an example of pose-based detection and tracking of tagged and untagged bees. Tags are detected in each frame, and the tag id is greedily assigned to a body detection based on a minimum distance below 50 pixels between the tag and the neck. Although the entire bodies were extracted for both datasets, as shown in Figure 4, the identity appearance models only used a cropped image of the abdomen to avoid any data leak: the tag image was is used only as the source of the groundtruth information and was removed completely from the data used in the study (see Figure 1E).

One challenge on bee identification is the abdomen curling that deforms the abdomen skin pattern, making the re-identification much more difficult. Other difficult samples exhibit occlusion and crowded images. Figure 2 show images from the raw dataset with examples of curling, occlusion, and crowded images. In this work, prospect instances where the individual has a curled abdomen were filtered out based on abdomen angle and size. Occluded bodies and crowded images were also excluded by removing the detections for which the distance from the waist keypoint to other bee waist was less than 300 pixels.

3.2 Short-term Dataset

The short-term dataset will be used solely for training purposes. The short-term dataset is based on tracklets of individuals. Each tracklet tracks the individual while its stays in the field of view of the camera. This data collection pipeline allows capturing a massive number of individuals in a short period. All the tracks were collected on the first 10 minutes of each hour from July 17 to July 24. All videos were downsampled to 10 fps. After the filtering based on abdomen size and angle, only tracks containing more than 10 images were kept.

This dataset contains 4,949 tracks; each track has an average of 22.15 images for a total of 109,654 images. On average, the length of a tracklet is about 5 seconds. The tracks were annotated with their track ID, meaning that different tracks are considered as different individuals for the triplet loss. We rely on the expectation that very few triplets will incorrectly select the same individual for the negative sample due to a large number of individuals. The examples in Figure 4a show that this dataset captures small variations mainly in the pose and illumination.

The short-term dataset was split randomly at track level into 80% training and 20% validation. It was not used for evaluation.

3.3 Long-term Dataset

The long-term dataset will be used both for training and evaluation purposes. The long-term dataset consists of marked bees that are monitored using barcode tags. This tagging allows monitoring the individuals on a long-term period. Due to physical limitations marking bees is limited to a few hundred individuals.

The dataset is split by July 17–23, July 24, and August 1– 4 for the training, validation, and testing set respectively, as shown in Figure 3. The long-term dataset ignores detections that do not have a tag associated. The training split contains 181 individuals on 801 tracks with a total of 3,777 images. The validation split contains 66 individuals on 309 tracks for a total of 1,909 images. The test split contains 126 individuals on 696 tracks for a total of 3,276 images. The test set has 29 identities that overlap the training set. The examples in Figure 4b show that this dataset captures drastic variations such as illumination, pose, and wings overlap.

4 METHOD

4.1 Embedding Network Architecture

The embedding network is a custom convolutional neural network (CNN) that takes the RGB image crop of the abdomen of the honeybee and outputs a 128 dimensional feature vector. Figure 5 shows the general architecture of the network. It consists of a 7x7

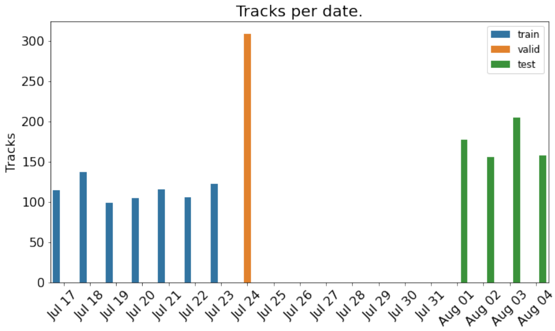

Figure 3: Number of tracks per day showing the split of the long-term dataset into training, validation and testing.

convolution layer with ReLU activation followed by 3 ResNet full pre-activation residual units (He et al., 2016), each with two 3×3 convolutions and output dimension of 64. The head of the network consists of a fully connected layer with an output of 128 dimensions. During training, dropout is applied before and after the fully connected layer with a probability of 0.5 and 0.2 respectively. The output of the network is L2 normalized.

Re-ID is performed by comparing the embedding of a query image to all embeddings of gallery images and ranking by Euclidean distance.

4.2 Training Protocols

Our training protocols consist of three modalities: 1) Fully Supervised (long-term), 2) Self-supervised (short-term), and 3) Supervised + Pretraining (short-term + long-term).

- The *fully supervised protocol* uses the long-term training dataset to minimize the distance of images of the same individual at different tracks.
- The *self-supervised protocol* uses the short-term dataset to minimize the distance of images on the same track. On the self-supervised protocol, the tracks were annotated with their track ID, meaning that different tracks are considered as different individuals for the triplet loss. We rely on the expectation that very few triplets will incorrectly select the same individual for the negative sample due to a large number of individuals.
- The *supervised + pretraining protocol* pre-trains the same as self-supervised protocol and finetunes the network with the fully supervised protocol.

The objective function for the three protocols is the Semi-Hard Triplet Loss with a margin of 0.2. Optimization is performed using Adam with a learning rate of 0.001 for 1000 epochs using early stop with a patience of 100 epochs. Data augmentation included: color distortion, color drop, gaussian blur and random cropping.

4.3 Evaluation Setups

The evaluation were always performed in the same way, independently from the training protocol, using only the test part of the long-term dataset, which provides ground-truth Re-ID information from the tags. It is based on a set of queries that are each compared to a gallery composed of one positive sample and 10 distractors. All of the queries and galleries are sampled randomly from the long-term test dataset under three scenarios of increasing difficulty: 1) same day, same hour; 2) different day, same hour of the day; 3) different day, any hour.

Pairs of tracks with the same ID were generated from all tracks on the long-term test dataset with additional conditions specific to each setup: 1) the same day-same hour protocol selects track pairs that are at least 15 minutes apart but not more than 60 minutes; 2) the diff day-same hour setup selects track pair that are on different days, but for which the time of the day is less than 60 minutes apart; 3) the diff day-any hour setup only enforces that the tracks are on different days. For each track in the pair, an image is randomly sampled from the track, and this process is repeated 100 times per track pair to generate an image query and its associated positive image sample. For each query, the 10 image distractors were sampled randomly from all negative IDs. The number of track pairs used for evaluation were 379, 236, and 1518 for the setups 1, 2, 3 respectively.

We report performance using the Cumulative Matching Characteristics (CMC) metric on rank-1 and rank-3, which represent the average rate at which a positive sample is ranked within the closest 1 (resp. 3) to its query.

5 RESULTS

In this section, we will evaluate the 3 training protocols in the 3 evaluation setups, and study the effect of both augmentation and amount of training data on the performance.

5.1 Base Performance

Table 2 shows that the short-term protocol outperforms the long-term protocol in all evaluation setups, by significant margins: +0.209 in same day, same hour, +0.188 in different day same hour, +0.136 in different day. This suggests that the short-term dataset

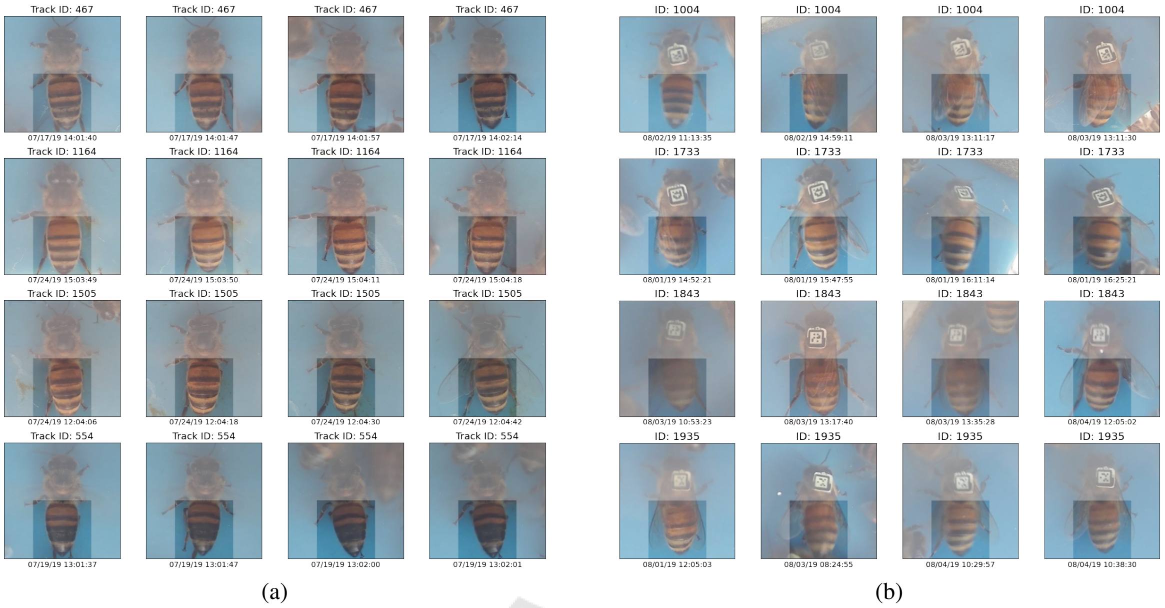

Figure 4: Example of pre-aligned images for the (a) Short-term and (b) Long-term datasets. The full image is shown to provide context, although only the highlighted abdomen area is used for training and evaluation, to ensure the tag that serves as groundtruth is not used by the model. Each row contains 4 images of the same individual.

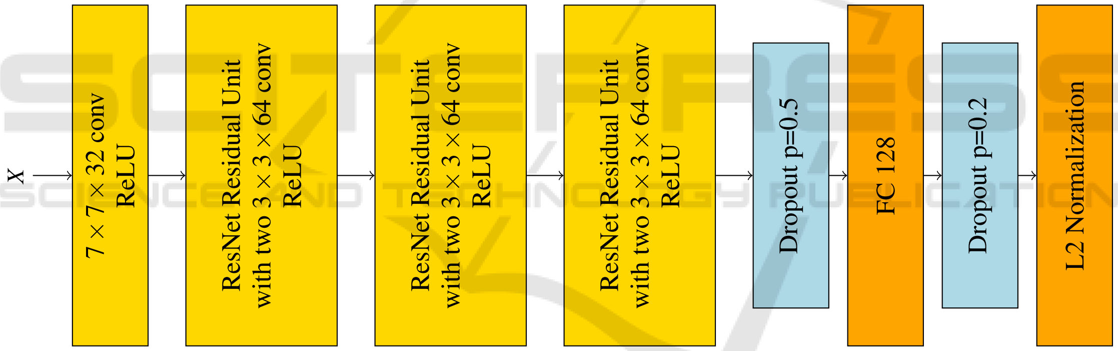

Figure 5: Model Architecture. The ResNet units are all full pre-activation residual units following 2×2 polling.

captures variations relevant for re-identification in all cases, although its advantage reduces with more challenging evaluation setups. Both approaches are outperformed by the protocol using short-term pre-training with long-term fine-tuning, which provides an improvement of: $+0.021$, $+0.072$ and $+0.088$ respectively over the short-term protocol itself in the approach with augmentation.

Table 2 shows that data augmentation improves the performance considerably in the short-term and the long-term training protocols. It has a consistent negative effect on the short-term + long-term training protocol.

5.2 Effect of Amount of Training Data

Although the long-term dataset has identities annotation over a long time span, previous sections showed it was at a disadvantage compared to the short-term dataset. We hypothesize a major factor is the lower amount of data. The long-term training data is significantly more challenging to obtain, as it requires extensive marking of individuals with barcode markers, which leads to a lower amount of unique individuals and lower amount of data usable for training. This section investigates how the amount of tracks affect the performance of the short-term training protocol to identify the trade-off between the quantity of data and its time scale.

Table 2: Cumulative Matching Characteristic performance of the three training protocols on the three evaluation setups. For each training protocol, performance is evaluated without and with augmentation, and the difference shown on the third row.

Method	**Training Protocol**	**Same day, same hour**		**Diff day, same hour**		**Diff day, any hour**	
		Rank 1	**Rank 3**	**Rank 1**	**Rank 3**	**Rank 1**	**Rank 3**
Triplet loss, No Aug	Long-term	0.456	0.733	0.322	0.610	0.273	0.557
Triplet loss, Aug		**0.529**	**0.781**	**0.391**	**0.709**	**0.362**	**0.664**
Triplet loss, No Aug	Short-term	0.682	0.888	0.508	0.775	0.436	0.720
Triplet loss, Aug		**0.738**	**0.912**	**0.579**	**0.831**	**0.498**	**0.788**
Triplet Loss, No Aug	Short-term	**0.801**	**0.932**	**0.659**	**0.889**	**0.624**	**0.868**
Triplet Loss, Aug	+ Long-term	0.759	0.913	0.651	0.880	0.586	0.832

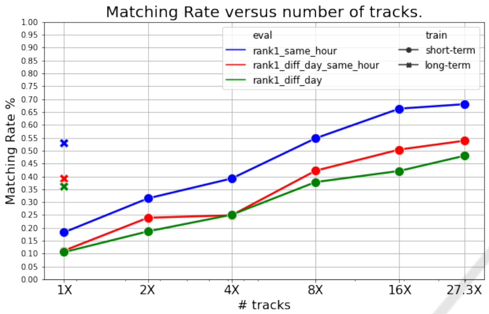

Figure 6: Effect of amount of training data on performance (CMC, rank 1). Marker shape represents the type of training (bullet for short-term vs crosses for long-term training data). Color represents the type of evaluation (blue for same day-same hour, red for diff day-same hour, green for diff day-any hour). Horizontal axis represents multiplicative factor from baseline of 181 tracks in short-term dataset, and 181 unique identities in long-term dataset.

For this experiment, the number of tracks in the short-term dataset was reduced to match the number of identities in the long-term dataset and create a baseline. As expected this baseline performs worst than long term, meaning that where the short-term and long-term have the same amount of identities the long-term dataset is much better. Figure 6 and Table 3 shows that the performance increases linearly with respect to the log of the number of tracks. When more data is available the performance increase up to the point where the short-term outperform the long-term as is shown in Figure 3. Due to the difficulty to gather more long-term training data, only the 1X factor is shown, which is limited by the number of marked bees. Collecting more short-term training data only requires processing more video in an unsupervised way. It should be investigated in future work at which point the performance increase from such additional data would start tapering off.

6 CONCLUSION

Animal re-identification is a challenging problem due to the lack of large-scale annotated datasets to both learn relevant models and evaluate performance in detail. In this paper, we proposed two main approaches to make progress. First, we contributed two large image datasets of honeybees. Both where extracted by leveraging automatic detection, pose estimation and tracking of honeybees from multiple days of video. All images have been compensated for position and orientation, as to provide well aligned images of honeybee bodies and their abdomen. The long-term dataset was annotated with the identity of 181 individuals recognized using barcode tags on their thorax, spanning up to 12 days. The second dataset was annotated based on short-term tracks IDs, which didn’t provide individual IDs, but could be collected at a larger scale than the long-term dataset (more than 12x the number of images).

The approach considered relies on contrastive learning with triplet-loss to train a 128 dimensions identity feature vector suitable for re-identification. The experimental study of the performance of re-identification showed the critical impact of the amount of data and the importance of data augmentation to maximize the performance. The results indicate that automated short-term tracking is a good approach to obtain the large amount of data required to learn re-identification models with limited human intervention. Although it cannot capture all possible degrees of variation a single individual may exhibit over long periods of time, it still capture relevant variations when enough data is collected, and ultimately outperforms training with the long-term dataset, which is more limited in terms of number of unique individuals due to the necessity of tagging physically the individuals to generate the groundtruth. Best performance was obtained by combining the two datasets, using the short-term data for pre-training and the long-term data for fine-tuning the Re-ID network.

Table 3: Effect of the amount of training data in the short-term training protocol with triplet loss and augmentation on CMC performance metric at rank-1 and rank-3.

# tracks	**Same day, same hour**		**Diff day, same hour**		**Diff day, any hour**	
	Rank 1	**Rank 3**	**Rank 1**	**Rank 3**	**Rank 1**	**Rank 3**
181	0.183	0.397	0.112	0.322	0.106	0.305
362	0.315	0.578	0.239	0.463	0.186	0.408
724	0.392	0.676	0.249	0.534	0.251	0.520
1448	0.548	0.795	0.422	0.688	0.378	0.655
2896	0.663	0.884	0.504	0.774	0.421	0.725
4949	**0.680**	**0.885**	**0.538**	**0.814**	**0.480**	**0.761**

These results show the possibility to recognize honeybees amongst a gallery of distractors over multiple days using only images of their abdomen. Future work will consider how the performance of such an approach would be improved with lightweight markings such as paint, by considering full-body images and by further increasing the scale of automatically collected training datasets, which could yield practical ways to track larger number of individuals over multiple hours and days without heavy marking procedures.

ACKNOWLEDGMENTS

This work is supported by grant no. 2021-67014-34999 from the USDA National Institute of Food and Agriculture. This material is based upon work supported by the National Science Foundation under grants no. 1707355 and 1633184. J. C. acknowledges support from the PR-LSAMP Bridge to the Doctorate, a program from the NSF under award number HRD-1906130. T. G. acknowledges NSF-HRD award #1736019 that provided funds for the purchase of bees. This work used the High-Performance Computing facility (HPCf) of the University of Puerto Rico, supported by National Institute of General Medical Sciences, National Institutes of Health (NIH) award number P20GM103475 and NSF grants number EPS-1002410 and EPS-1010094. This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number ACI-1548562. Specifically, it used the Bridges-2 system, which is supported by NSF award number ACI-1928147, at the Pittsburgh Supercomputing Center (PSC).

REFERENCES

Bergamini, L., Porrello, A., Dondona, A. C., Negro, E. D., Mattioli, M., D'alterio, N., and Calderara, S. (2018). Multi-views Embedding for Cattle Re-identification. In *2018 14th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*, pages 184–191.

Boenisch, F., Rosemann, B., Wild, B., Dormagen, D., Wario, F., and Landgraf, T. (2018). Tracking All Members of a Honey Bee Colony Over Their Lifetime Using Learned Models of Correspondence. *Frontiers in Robotics and AI*, 5:35.

Bozek, K., Hebert, L., Portugal, Y., and Stephens, G. J. (2021). Markerless tracking of an entire honey bee colony. *Nature Communications*, 12(1):1733.

Brust, C.-A., Burghardt, T., Groenenberg, M., Kading, C., Kuhl, H. S., Manguette, M. L., and Denzler, J. (2017). Towards automated visual monitoring of individual gorillas in the wild. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 2820–2830.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. (2020a). A simple framework for contrastive learning of visual representations. *Proceedings of the 37th International Conference on Machine Learning*, 119:1597–1607.

Chen, T., Kornblith, S., Swersky, K., Norouzi, M., and Hinton, G. E. (2020b). Big Self-Supervised Models are Strong Semi-Supervised Learners. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 22243–22255. Curran Associates, Inc.

Deb, D., Wiper, S., Gong, S., Shi, Y., Tymoszek, C., Fletcher, A., and Jain, A. K. (2018). Face recognition: Primates in the wild. In *2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pages 1–10. IEEE.

Gao, J., Burghardt, T., Andrew, W., Dowsey, A. W., and Campbell, N. W. (2021). Towards Self-Supervision for Video Identification of Individual Holstein-Friesian Cattle: The Cows2021 Dataset. *arXiv preprint arXiv:2105.01938.*

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Identity Mappings in Deep Residual Networks BT - Computer Vision – ECCV 2016. pages 630–645, Cham. Springer International Publishing.

Körschens, M., Barz, B., and Denzler, J. (2018). Towards automatic identification of elephants in the wild. *arXiv preprint arXiv:1812.04418.*

Li, S., Li, J., Lin, W., and Tang, H. (2019). Amur tiger re-identification in the wild. *arXiv preprint arXiv:1906.05586.*

Li, W., Zhao, R., Xiao, T., and Wang, X. (2014). DeepReID: Deep Filter Pairing Neural Network for Person Re-identification. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 152–159.

Mégret, R., Rodriguez, I. F., Ford, I. C., Acuña, E., Agosto-Rivera, J. L., and Giray, T. (2019). LabelBee: A Web Platform for Large-Scale Semi-Automated Analysis of Honeybee Behavior from Video. In *Proceedings of the Conference on Artificial Intelligence for Data Discovery and Reuse*, AIDR '19, New York, NY, USA. Association for Computing Machinery.

Misra, I. and Maaten, L. v. d. (2020). Self-supervised learning of pretext-invariant representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6707–6717.

Noroozi, M. and Favaro, P. (2016). Unsupervised learning of visual representations by solving jigsaw puzzles. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 9910 LNCS, pages 69–84.

Pereira, T. D., Tabris, N., Li, J., Ravindranath, S., Papadoyannis, E. S., Wang, Z. Y., Turner, D. M., McKenzie-Smith, G., Kocher, S. D., Falkner, A. L., Shaevitz, J. W., and Murthy, M. (2020). SLEAP: Multi-animal pose tracking. *bioRxiv*, page 2020.08.31.276246.

Rodriguez, I. F., Mégret, R., Egnor, R., Branson, K., Agosto, J., Giray, T., and Acuna, E. (2018). Multiple animals tracking in video using part affinity fields. In *Workshop on visual observation and analysis of vertebrate and insect behavior. In Proceedings of the 24th International Conference on Pattern Recognition (ICPR), Beijing, China*, pages 20–24.

Romero-Ferrero, F., Bergomi, M. G., Hinz, R. C., Heras, F. J. H., and de Polavieja, G. G. (2019). idtracker.ai: tracking all individuals in small or large collectives of unmarked animals. *Nature Methods*, 16(2):179–182.

Schneider, S., Taylor, G. W., and Kremer, S. C. (2020). Similarity learning networks for animal individual re-identification-beyond the capabilities of a human observer. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision Workshops*, pages 44–52.

Schofield, D., Nagrani, A., Zisserman, A., Hayashi, M., Matsuzawa, T., Biro, D., and Carvalho, S. (2019). Chimpanzee face recognition from videos in the wild using deep learning. *Science Advances*, 5(9):eaaw0736.

UN Press material (2019). UN Report: Nature's Dangerous Decline 'Unprecedented'; Species Extinction Rates 'Accelerating'. *UN Sustainable Development Goals*, page 19.

Wang, J. and Olson, E. (2016). AprilTag 2: Efficient and robust fiducial detection. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4193–4198.

Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., and Tian, Q. (2015). Scalable Person Re-identification: A Benchmark. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1116–1124.