Low-power, Continuous Remote Behavioral Localization with Event Cameras

Friedhelm Hamann[1,5], Suman Ghosh[1], Ignacio Juárez Martínez[2],
Tom Hart[3], Alex Kacelnik[2,5] and Guillermo Gallego[1,4,5].
[1] Technische Universität Berlin, [2] Oxford University, [3] Oxford Brookes University,
[4] Einstein Center for Digital Future, [5] Science of Intelligence Excellence Cluster.

Supplementary Material

Section 7.1 provides more information on the biological motivation of our project. Section 7.2 shows per nest results of our method. In Sec. 7.3, we report the number of proposals and training run time. Section 7.4 gives more details about data acquisition, filtering, and split. Additionally, Sec. 7.5 provides several sensitivity studies. Lastly, Sec. 7.6 shows results for a naive baseline for the random classifier.

7.1. Biological Motivation and Impact

"Displays" are stereotyped sequences of movements that are key to communication between animals of the same species. In penguins, these behaviors are widespread and are used for a variety of purposes including mate choice and pair bonding. These displays are accompanied by vocalizations that are known to be individually distinctive and to allow both mate and chick recognition in most species [1].

In this paper, we choose to study the ecstatic display (ED), one of the most common and recognizable displays in *Pygoscelid* penguins. During the ecstatic display penguins stand fully erect on their nests with their stretched neck and bill pointing up vertically. They move their outstretched flippers back and forth in fast beats while they emit very loud rasps that make their chest vibrate synchronously [1,2]. These rasps are normally emitted in pairs of syllables made up of a short inhale followed by a long and loud exhale [3]. Here we study this behavior in Chinstrap penguins (*Pygoscelis antarctica*) for which there is an almost complete lack of information regarding their displays [2].

Studies in the two closest species (*Pygoscelis adeliae* and *Pygoscelis papua*) suggest the ecstatic display could be an "honest display" intended for males to communicate body conditions to females and/or defend the nesting area from nearby males without a fight [4, 5]. This hypothesis arises because, in those two species, ED occurs only in males at the beginning of the season [2], when they claim a nest and compete to attract a partner. In chinstrap penguins, however, this behavior happens throughout the season and is displayed by both sexes (as we see through our event camera recordings). We understand that this behavior could serve a different communication purpose in this species and we want to explore whether it indeed mediates similarly important functions for pair formation and colony structuring as in the other two species. To find out, first, we must understand when, how, and how long this behavior occurs before drawing relationships to other factors like sex, breeding stage, at-sea behavior, or environmental factors. Behavioral monitoring like this is proving increasingly important to anticipate changes in breeding habits before a population decline occurs [6]. With this work, we aim to open the door for other researchers to use event cameras and TAD for "large-scale" automatic detection of behaviors in preventive monitoring.

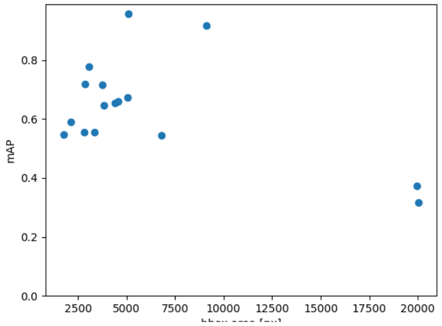

Figure 1. Visualization showing the relation between bounding box size and mAP.

7.2. Results per nest

The best results of our method per penguin nest are shown in Tab. 1 and visualized in Fig. 1. While it is difficult to extract trends from different nests, the results show a large variation in the data. It furthermore hints at the importance of good camera positioning during data acquisition. An elevated camera position, which allows clear separation of the nests, aids in the accurate positioning of the bounding boxes. This is a lesson learned for future data acquisition campaigns. We expect to be able to improve the usability of the proposed method by using more cameras and recording

mAP@IoU	N01	N02	N03	N04	N05	N06	N07	N08	N09	N10	N11	N12	N13	N14	N15	N16
0.1	0.53	0.45	0.66	1.00	0.67	0.67	0.85	0.76	0.76	0.98	0.72	0.86	0.55	0.72	0.67	0.69
0.3	0.40	0.45	0.66	1.00	0.67	0.67	0.85	0.69	0.73	0.98	0.72	0.86	0.55	0.72	0.65	0.67
0.5	0.30	0.33	0.62	1.00	0.67	0.44	0.58	0.69	0.73	0.98	0.71	0.79	0.55	0.72	0.53	0.59
0.7	0.27	0.03	0.25	0.67	0.67	0.44	0.35	0.47	0.66	0.90	0.43	0.59	0.55	0.72	0.38	0.40
Average	0.37	0.32	0.55	0.92	0.67	0.56	0.66	0.65	0.72	0.96	0.65	0.78	0.55	0.72	0.55	0.59
# ED	8.00	6.00	11.00	3.00	9.00	3.00	4.00	8.00	11.00	6.00	15.00	13.00	7.00	6.00	21.00	10.00

Table 1. Results per nest, which show the variation with respect to the data.

Proposal Method	# proposals	AR, Top 50
Sliding Window	12820320	0.08
Watershed	352	0.27
event TAG [7]	13117	0.49
reTAG (Ours)	30527	**0.66**

Table 2. Number of generated proposals for the test set and the average recall (AR) for different methods.

fewer nests per camera.

7.3. Runtime, Computational Effort

The number of generated proposals for the test set per method is reported in Tab. 2. We can see the advantage of the TAG-based methods compared to the sliding window approach regarding the number of proposals. The watershed algorithm is too simple and does not produce a sufficient amount of proposals. Comparing the two TAG-based algorithms shows the trade-off between recall and computational effort. While our reTAG has significantly increased recall, it also generated five times more proposals. Overall, our reTAG is lightweight compared to the sliding window approach and has 8× higher average recall.

Inference time: The ATSN with a ResNet18 backbone has a run time of 2.56ms (Nvidia Tesla V100S) for one example and one forward pass. Each time interval generated in the first step (proposal) is a sample for training and validation of the second step (classifier).

Training time: For the training set the reTAG algorithm outputs 189519 proposals. To maintain a manageable balance between foreground (ED) and background, the negative samples in the training set are sub-sampled by a factor of 10, leading to a training set with 2093 positive and 18859 negative samples. We train for 10 epochs, resulting in a training time for the classifier of approximately 30 minutes on an Nvidia Tesla V100S GPU.

7.4. Dataset Details

In total, we collected around 238 hours of unfiltered data from the Vapour Col penguin colony in Deception Island, in the form of ROSBag files. The ROSBag files were then post-processed using a hot pixel filter [8], which discards data from faulty pixels that trigger events at a high rate. Among the recorded data, we selected 24 ten-minute-long sequences for annotation of ecstatic display (ED) behavior. The selected sequences include diverse scenarios from different dates and hours of the day, to account for various illumination and weather conditions.

Incidentally, our setup was deployed next to a foraging camera which takes a snapshot of the penguin colony every 1 minute. During the night, the foraging camera uses flashes of infrared (IR) light to acquire images in low light. Since the event camera sensor is also sensitive to IR light, the aforementioned flashes produce a flurry of events throughout the scene once per minute. Consequently, events generated by IR flashes were filtered out from the night sequences before applying our proposed TAD algorithm.

Detailed information on every ten-minute sequence in the annotated dataset can be found in Tab. 3. The table furthermore indicates the data split (training, validation, testing) we used in all experiments. The test split reflects the proportion of precipitation in Antarctica during the breeding season.

7.5. Additional Sensitivity Studies

Table 4 reports the results using a MobileNetV3 backbone instead of ResNet18. The figures are similar with both backbone variants and event representations (only a 4% performance drop on average), supporting the robustness of our method to different design choices.

Similarly, Tab. 5 shows results concerning different values of the accumulation time Δt for histograms and decay τ for time maps.

7.6. Naive Baseline Classifier

There is a high class imbalance in the proposals. Randomly guessing leads to a high number of false positives, and consequently a low precision. To confirm this, we implemented a random classifier (akin to flipping a coin), which accepts a proposal with a 50% change, setting the score to 1 for mAP calculation. This solution achieves 0.035% mAP (Avg.).

Day	time	split	night	precipitation	#ED
Jan 5th	17:00	train	✗	✗	2
Jan 6th	19:00	train	✗	✗	11
Jan 7th	05:00	train	✗	✓	70
Jan 7th	08:00	train	✗	✗	3
Jan 9th	20:04	train	✗	✗	28
Jan 11th	21:06	train	✗	✗	9
Jan 12th	03:36	train	✓	✓	54
Jan 12th	03:56	train	✗	✓	23
Jan 12th	08:56	train	✗	✓	0
Jan 12th	12:56	train	✗	✗	0
Jan 12th	17:26	train	✗	✗	58
Jan 13th	00:00	train	✓	✗	92
Jan 13th	10:59	train	✗	✗	3
Jan 13th	14:59	train	✗	✓	11
Jan 14th	23:58	train	✓	✗	1
Jan 15th	13:58	train	✗	✗	0
Jan 18th	02:56	train	✓	✗	0
Jan 7th	02:00	validation	✓	✗	20
Jan 17th	15:56	validation	✗	✗	29
Jan 6th	01:00	test	✓	✗	53
Jan 13th	09:59	test	✗	✗	47
Jan 14th	21:58	test	✗	✗	8
Jan 15th	05:58	test	✗	✗	18
Jan 15th	11:48	test	✗	✓	25

Table 3. An overview of all ten-minute sequences in the annotated dataset.

Backbone	Event repres.	0.1	0.3	0.5	0.7	Average
ResNet18 (Tab. 3)	Time-map	**0.66**	**0.64**	**0.58**	**0.43**	**0.58**
MobileNetV3-Large	Time-map	0.62	0.59	0.53	0.36	0.53
MobileNetV3-Large	Histogram	0.57	0.54	0.49	0.36	0.49
MobileNetV3-Small	Time-map	0.60	0.56	0.50	0.35	0.50
MobileNetV3-Small	Histogram	0.50	0.48	0.45	0.32	0.44

Table 4. Sensitivity of the system with respect to the backbone (ResNet or MobileNet) and input represesntation. Mean Average Precision at several IoU levels (mAP@IoU).

	Histogram Δt				Time map: Decay τ [s]		
Δt [s]	0.3	1	3	τ [s]	0.02	0.2	2
mAP	0.55	0.56	0.52	mAP	0.45	0.58	0.51

Table 5. Sensitivity of the system concerning parameters of the frame representation. Mean average precision at several IoU levels (mAP@IoU).

References

[1] Pierre Jouventin and F Stephen Dobson, *Why penguins communicate: the evolution of visual and vocal signals*. Academic Press, 2017.

[2] Tony D. Williams, *The Penguins: Spheniscidae*. Oxford: Oxford University Press, 1995.

[3] Javier Bustamante and Rafael Márquez, "Vocalizations of the chinstrap penguin pygoscelis antarctica," *Colonial Waterbirds*, pp. 101–110, 1996.

[4] Maureen A Lynch and Heather J Lynch, "Variation in the ecstatic display call of the gentoo penguin (pygoscelis papua) across regional geographic scales," *The Auk: Ornithological Advances*, vol. 134, no. 4, pp. 894–902, 2017.

[5] Emma Marks, Allen Rodrigo, and Dianne Brunton, "Ecstatic display calls of the adélie penguin honestly predict male condition and breeding success," *Behaviour*, vol. 147, no. 2, pp. 165–184, 2010.

[6] Francesco Cerini, Dylan Z Childs, and Christopher F Clements, "A predictive timeline of wildlife population collapse," *Nature Ecology & Evolution*, pp. 1–12, 2023.

[7] Guang Chen, Sanqing Qu, Zhijun Li, Haitao Zhu, Jiaxuan Dong, Min Liu, and Jörg Conradt, "Neuromorphic vision-based fall localization in event streams with temporal–spatial attention weighted network," *IEEE Trans. Cybern.*, vol. 52, no. 9, pp. 9251–9262, 2022.

[8] Cedric Scheerlink, "DVS hot pixel filter." https://github.com/cedric-scheerlinck/dvs_tools/tree/master/dvs_hot_pixel_filter, 2019.