Animal Behaviour 177 (2021) 31–37

Contents lists available at ScienceDirect

Animal Behaviour

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

Commentary

Automating the analysis of fish grazing behaviour from videos using image classification and optical flow

Ellen M. Ditria[*], Eric L. Jinks, Rod M. Connolly

Coastal and Marine Research Centre, Australian Rivers Institute, School of Environment and Science, Griffith University, Gold Coast, QLD, Australia

ARTICLE INFO

Article history:
Received 6 May 2020
Initial acceptance 20 July 2020
Final acceptance 9 March 2021

MS. number: 20-00333R

Keywords:
animal behaviour
aquatic
automation
deep learning
image classification
machine learning
optical flow

Studying and quantifying behaviour is important to understand how animals interact with their environments. However, manually extracting and analysing behavioural data from the large volume of camera footage collected is often time consuming. Deep learning techniques have emerged as useful tools in automating the analysis of certain behaviours under controlled or laboratory conditions, but the complexities of using raw footage from the field has resulted in this technology remaining largely unexplored as a possible data analysis alternative for animals in situ. Here, we use deep learning techniques to automate the analysis of fish grazing behaviour from real-world field imagery. We collected video footage in sea grass meadows in Queensland, Australia, and trained models on a training data set of over 3000 annotations. We used a combination of dense optical flow to assess pixel movement in underwater footage, spatiotemporal filtering to increase accuracy, and deep learning algorithms to classify grazing behaviour of luderick, *Girella tricuspidata*. When tested on novel videos the model had not seen in training, the model correctly identified nearly all individual grazing events. Deep learning shows promise as a viable tool for determining animal behaviour from underwater videos, and with further development offers an alternative to current time-consuming manual methods of data extraction.

The way researchers study animal behaviour has evolved rapidly over the last few decades. New methods are proving to be important in understanding complex biological processes, and recent technological advances have provided techniques to solve the challenges of traditional data collection and analytical methods (Hughey, Hein, Strandburg-Peshkin, & Jensen, 2018). Historically, trained scientists would directly study and quantify animal behaviour in situ, limiting the number of direct observations that could be made as well as the spatial and temporal resolution of the observations (Altmann, 1974). More recently, techniques and equipment such as acoustic telemetry, accelerometers and GPS tags have become popular in monitoring animal behaviour (Espinoza, Farrugia, Webber, Smith, & Lowe, 2011; Ladds et al., 2017; Browning et al., 2018). However, these devices require manual and invasive application techniques that also limit sample size due to cost and logistics. Alternatively, cameras are widely used to obtain behavioural data and capture information both on animals and their interactions with the environment. Camera techniques allow the study of these behaviours and interactions on a fine spatial scale without the need for invasive procedures (Hughey et al., 2018). Despite this, the use of camera techniques in quantitative behavioural studies is often hampered by time-consuming manual analysis of animal behaviour data from video footage (Han, Taralova, Dupre, & Yuste, 2018; Weinstein, 2017).

Deep learning techniques are emerging as a useful solution to assist with the data analysis bottleneck often encountered by manual methods of video analysis. Deep learning is a subcategory of machine learning that has begun to emerge in the environmental sciences due to its ability to automatically process raw videos and images, data that historically needed to be analysed manually to extract information. Deep learning frameworks consist of a number of computational layers that can process raw data images and automatically extract features at the pixel level, unlike traditional machine learning algorithms, such as support vector machines, which require human input to manually extract features for the algorithm to recognize (LeCun, Bengio, & Hinton, 2015). Deep learning methods can perform faster than humans and are often equal to or more accurate than manual analysis (Villon et al., 2018; Torney et al., 2019). Given the advantages of deep learning, the use of these methods for analysing video data is beginning to be applied in animal behaviour. Several studies have been able to detect the presence of animals and their behaviours and activity using acoustic data recordings (Strout et al., 2017; Himawan, Towsey, Law, & Roe, 2018; Xie & Zhu, 2019). However, for

[*] Corresponding author.
E-mail address: ellen.ditria@griffithuni.edu.au (E. M. Ditria).

https://doi.org/10.1016/j.anbehav.2021.04.018
0003-3472/© 2021 The Association for the Study of Animal Behaviour. Published by Elsevier Ltd. All rights reserved.

nonvocal species, this approach is not suitable. Deep learning algorithms have also been implemented to extract behaviour information from video and images, although largely through efforts in highly controlled environments (Valletta, Torney, Kings, Thornton, & Madden, 2017; Yang et al., 2018; Xu, Bennamoun, An, Sohel, & Boussaid, 2019). Field data present substantial challenges for automating the analysis of behaviour relative to standardized laboratory tests. Animals often appear at different angles and distances to the camera, or are only partially visible, making not only behaviour difficult to determine but, in some cases, also the detection of the animal itself (Norouzzadeh et al., 2018; Sun et al., 2018; Nguyen et al., 2019). Deep learning models must be adequately trained on these different scenarios to perform adequately, which can require a substantial amount of training data. However, deep learning techniques have proven to be well suited to inferring information from objects of interest varying in size and shape, even if partially obscured, so they do have the potential for behavioural data analysis in field studies (Long, Shelhamer, & Darrell, 2015).

Dense optical flow is a method of analysing the directional movement of all pixels between frames of video footage (Farnebäck, 2003). Optical flow algorithms have been used to assess movement since the 1980s (Gultekin & Saranli, 2013; Walker, Gupta, & Hebert, 2015), but have only recently been coupled with deep learning algorithms to more accurately analyse movement from videos. Deep learning and dense optical flow have proven to be complementary technological tools in assessing movements such as sow nursing behaviour in agricultural livestock (Yang et al., 2018). However, the combination has rarely been used for wild animals. In one of the few examples, Golkarnarenji et al. (2018) used dense optical flow and deep learning algorithms to identify the Baw Baw frog, *Philoria frosti*, by identifying moving pixels in frames and separating them from still, empty frames in large volumes of data from camera trap footage. This study pointed to the advantages of using deep learning algorithms by showcasing their ability to automatically remove vast numbers of empty frames and thus dramatically reduce manual processing time.

Utilizing dense optical flow and deep learning algorithms in aquatic environments presents a unique set of challenges. Although deep learning has been used for identifying and counting aquatic species from above-water footage (Jovanović, Svendsen, Risojević, & Babić, 2018; Gray, Bierlich et al., 2019; Gray, Fleishman et al., 2019; Zhou et al., 2019), surface behaviours constitute only a subset of behaviours exhibited by many aquatic animals. Fish behaviour in particular is relatively underrepresented in the literature compared with other vertebrates, despite fish having twice the number of species of birds and mammals combined (Rosenthal, Gertler, Hamilton, Prasad, & Andrade, 2017). This is likely to be due to logistical difficulties and poor-quality footage produced from collecting and analysing data in underwater environments (Sun et al., 2018). To avoid these environmental difficulties, most footage taken for deep learning analysis in aquatic ecosystems has primarily been attempted in coral reef habitats, which have environmental advantages compared to other aquatic habitats, such as high visibility and light (Xu & Matzner, 2018).

The species of interest in this case study is a common herbivorous fish in coastal waters in eastern Australia, the luderick, *Girella tricuspidata*. This species displays an identifiable sideways sweeping motion when stripping epiphytic algae from blades of sea grass. Using this feeding behaviour as a target, we determined whether the combination of deep learning and dense optical flow techniques could distinguish between algal grazing and nongrazing (swimming) behaviours in this species. In doing so, we developed automated behavioural analysis of animal feeding in aquatic environments, something not previously attempted from in situ underwater videos.

METHODS

Training Data Set

Free-swimming luderick were recorded using submerged action cameras (Haldex Sports Action Cam HD 1080p) in the Tweed River estuary on the border of Queensland and New South Wales, Australia (-28.169438S, 153.547594E), between February and July 2019. Each sampling day, six cameras were deployed for 1 h over a variety of sea grass patches with varying angle and camera placement to ensure a variety of backgrounds and fish angles were captured. Videos were manually trimmed for training to contain footage of luderick and were split into 25 frames/s for manual annotation of still images. Segmentation polygons were drawn around individual luderick and were annotated as either grazing or nongrazing. The training data set was balanced to include approximately an equal number of both behaviours to avoid prediction bias from having disproportionate training data for one particular behaviour. This is important as training only on a single behaviour would increase the likelihood of that behaviour being predicted by the algorithm, resulting in a greater number of false positives. Algal grazing behaviour was defined from the point when a fish placed its mouth on a sea grass blade until it left the blade. Our resulting training data set consisted of 3396 annotations which took a total of 7.25 h to label (approximately eight annotations manually generated per minute).

Convolutional Neural Network and Dense Optical Flow

The deep learning algorithm used is called ResNet50, a widely used framework developed for image classification (He, Zhang, Ren, & Sun, 2016), which was pretrained on the ImageNet-1k data set (http://www.image-net.org). Model training, testing and prediction tasks were conducted on a Microsoft Azure Data Science Virtual Machine powered by an NVIDIA V100 GPU via a desktop computer. Models took a total of approximately 3.5 h to reach 40 000 training iterations (where the model has ‘studied’ the whole training data set 40 000 times). Optical flow data were generated from the raw video images based on the relative movement of pixels between frames using a motion estimation algorithm (Farnebäck, 2003). Motion estimation was conducted across two adjoining frames for each annotated polygon (i.e. luderick), and output as a red-green-blue (RGB) image (Fig. 1; Farnebäck, 2003). Each polygon from the optical flow output image was used to train the deep learning model. This model classified the behaviour within the polygon of the luderick. Three models were trained and tested on the same data sets to account for variation in the training data due to randomized data augmentation. A standard data augmentation technique was employed, random horizontal flipping, where random frames from 50% of the data set are chosen and inverted to increase the volume of training data.

Spatiotemporal Filtering and Data Interpolation

Grazing behaviour occurs over several frames of video footage, yet our metrics to quantify accuracy are reliant on predictions per frame. Model accuracy can be enhanced when knowledge of previous and subsequent frames is considered by the algorithm. To achieve this, we applied spatiotemporal filtering (STF) to remove fleeting temporal anomalies in the model's predictions. If the model incorrectly identifies one frame as demonstrating grazing behaviour but the previous or subsequent frames do not, the frame was removed as a positive prediction. Similarly, data were interpolated for single frames predicted to be not grazing that did exhibit grazing behaviour in previous and subsequent frames. Each

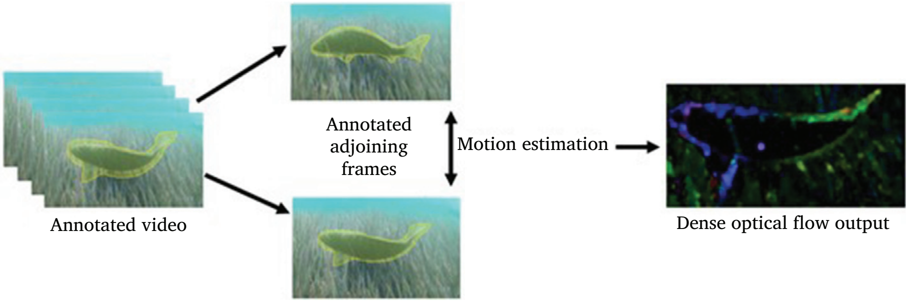

Figure 1. Measuring movement using optical flow. Raw videos are split into 25 frames/s and the luderick annotated. Annotated frames are run through the optical flow algorithm, which estimates motion between adjoining frames. Pixel colour in output images denotes pixel movement between frames; black pixels indicate no movement.

of the models was run both with and without STF to determine whether this additional step increased model accuracy.

Test Set and Performance Metrics

A data set of 18 videos comprising unseen footage from Tweed Estuary of both grazing and nongrazing behaviours was annotated as the ground-truthed test data. Performance was evaluated per frame and per video. Per frame, performance was calculated by the F1 score; the harmonic mean of recall and precision (Everingham, Van Gool, Williams, Winn, & Zisserman, 2010). Precision is how rigorous the model is at identifying the presence of luderick grazing behaviour, and recall is the fraction of the total positives the model correctly classified (Everingham et al., 2010). The F1 score determines how well the model can identify the behaviour by combining recall and precision with equal weighting. The performance metrics were calculated as follows:

$$\text{Precision }(P) = \frac{\text{True positive}}{\text{True positive} + \text{False positive}}$$

$$\text{Recall }(R) = \frac{\text{True positive}}{\text{True positive} + \text{False negative}}$$

$$\text{F1 score} = 2 \times \frac{P \times R}{P + R}$$

Per video, performance was measured as the number of grazing events detected against the manually ground-truthed number of events.

Statistical Analysis

A mixed-effects linear model was run to determine whether the use of STF influenced model performance (recall, precision and F1 score). The fixed effect was STF and the random effect was the individual model.

All data labelling, training and testing were conducted on FishID, a cloud-based software developed at Griffith University (https://globalwetlandsproject.org/tools/fishid/). The code used for the deep learning, optical flow and spatiotemporal filtering is provided in the Supplementary material.

RESULTS

The method proved effective at detecting grazing behaviour. On a per frame basis, the vast majority of frames were correctly predicted as grazing and nongrazing, with high numbers of true positives, relatively few false negatives, and fewer again false positives (Fig. 2a). Without spatiotemporal filtering (STF), recall, precision and F1 were all between 73 and 79% (Fig. 2b). STF improved performance substantially, increasing the proportion of true positives. The addition of STF significantly increased recall ($F_{1,2} = 52$, $P<0.001$), precision ($F_{1,2} = 24$, $P = 0.039$) and F1 ($F_{1,2} = 72$, $P = 0.01$) to between 84 and 87% (Fig. 2b). The method mostly reported false negatives as fleeting mistakes (e.g. missing the grazing behaviour when the fish was obscured momentarily by sea grass while feeding). False positives sometimes occurred over longer periods, with the model mistaking another behaviour (swimming) for grazing (see Appendix Table A1). Per video, the performance of the method on a per video basis was very good. In all, 34 of the 37 grazing events (92%) were correctly detected when STF was applied (Table 1). The number of events per video was generally the same as the manually ground-truthed number, or nearly so, for all 18 test videos (number of events ranging from 1 to 7). The number of grazing events was substantially less without the application of the STF, only 27 of 37 events. Example videos of a trained model detecting grazing behaviour can be found in the Supplementary material.

DISCUSSION

We have provided early evidence that successful behavioural classification of grazing movements using the combination of optical flow and deep learning techniques from unconstrained underwater environments may be a viable technique. We have shown that these combined techniques can identify grazing behaviour of segmented luderick in underwater videos taken in situ in marine environments. We have also shown that spatiotemporal filtering (STF) is a useful postprocessing step for obtaining higher accuracy. By learning optical flow patterns representing specific body movements, our models correctly recorded 92% of grazing events in the test videos.

While the deep learning method proposed in this paper can identify behaviour with relatively high accuracy, detecting errors within the system is still difficult to control. Although deep learning algorithms have been shown to have a lower error rate than human counterparts when tested on data sets similar to ones they have been trained on (Ditriaa et al. 2020a), some form of error or bias is present in all methods of data analysis; for deep learning, this is especially the case when presented with novel data different to what the algorithm has been trained on. Methods for controlling error rates for deep learning classification algorithms are beginning to emerge. For example, Villon et al. (2020) proposed a framework that classifies multiple fish species with not only a species name, but also an additional postprocessing label of ‘sure’ or ‘unsure’ based on a set confidence threshold after the species is determined by the deep learning algorithms. While this may still require manual effort to determine the species in the ‘unsure’ category, the

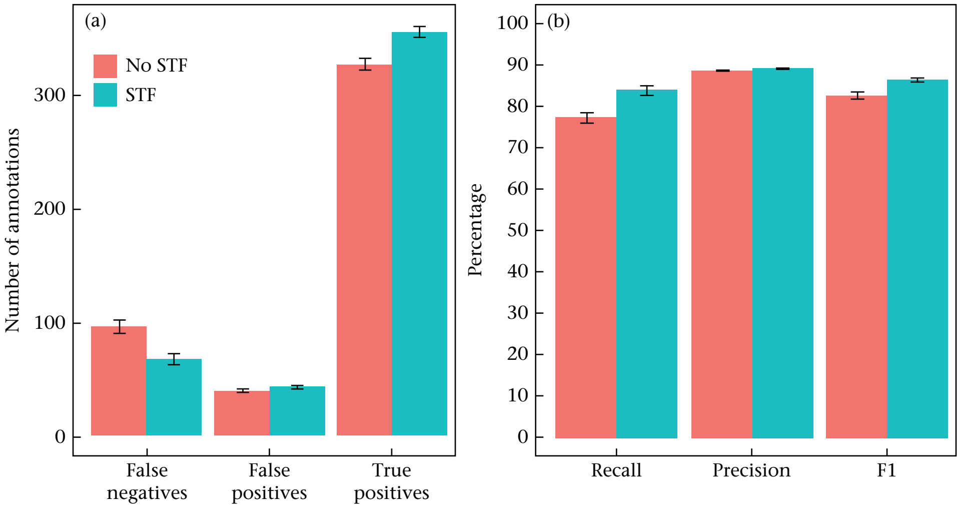

Figure 2. Performance of automated behaviour models on a per frame basis, with and without spatiotemporal filtering (STF). (a) The number of frames classified correctly (True positives) and falsely (False negatives, False positives). (b) Recall, precision and F1 scores. $N = 3$ models in all cases. Vertical lines represent SEs.

Table 1
Model performance measured as detection of grazing events

Video	Manual ground-truth (no. of events)	Computer detection with STF (no. of events)	Computer detection without STF (no. of events)
1	1	1.0	1.0
2	7	7.0	5.0
3	1	0.3	0.0
4	4	3.3	2.7
5	1	1.0	1.0
6	4	4.0	3.0
7	1	1.0	0.3
8	1	1.0	0.3
9	2	2.0	0.7
10	4	2.6	2.3
11	1	1.0	1.0
12	2	2.0	2.0
13	1	1.0	1.0
14	2	2.0	2.0
15	1	1.0	1.0
16	1	1.0	1.0
17	1	1.0	1.0
18	2	2.0	1.0
Overall	37	34.2	27.3

Results are presented for individual test videos, and overall, with and without spatiotemporal filtering. No. of events for manual ground-truthing are exact values. No. of events for computer detections are an average of three models.

overall workload is significantly lessened, and the misclassification error rate was reduced from 22% to 3%. This type of statistical approach to deep learning outputs might also be effective in behavioural research.

The automation method we describe can be a complementary tool to research that already relies on camera technology to capture behaviours that are usually analysed manually. Although our models were successful at identifying grazing versus nongrazing behaviours, there are some important caveats and biases with utilizing deep learning and optical flow to automate this process. Individuals of our target species were of an easily observable size (range approximately 30–40 cm long); they are common in coastal habitats, and display an obvious body movement when grazing, making them a good model species to test this method. The method might be less suitable for smaller or cryptic species, or those that display subtle behaviours. However, given that we suggest these techniques can replace current manual analysis of behaviours from video footage, it is likely that only perceptible behaviours are documented to begin with. Researchers will need to address whether to implement automation of video analysis on a case-by-case basis to decide whether it is appropriate for the specific behaviour and species of interest.

Deep learning provides a cost-effective way to analyse large volumes of video data. In Ditria et al. (2020a) we showed that once trained, a deep learning algorithm could identify and count fish in isolated images and segments of video not only faster than humans, but also more accurately and more consistently. However, the application of automated methods often requires higher start-up costs and effort, followed by lower costs after initial implementation. For example, González-Rivero et al. (2020) showed that automatic processing of coral reef images using deep learning resulted in a 99% cost reduction and was 200 times faster than manual methods. However, this may not be a suitable method for all studies. If the research project consists only of a small amount of data to be processed as a singular study, manual processing may prove to be most cost effective. Here, we have provided the time it took to label the data set and train the model as a guide; times will be project specific and will vary depending on computation resource and manual skills in identifying behaviours. The length of the study and the amount of data to be processed should be taken into consideration when deciding whether deep learning methods are a suitable approach for behavioural studies.

Not only can deep learning techniques recognize animals at the species level, they can also identify individuals from a population (Konovalov et al., 2018; Hou et al., 2020), suggesting that this technique can be implemented to quantify individual differences in feeding behaviours. This technique could be applied to ascertain the frequency of certain behaviours influenced by a number of biotic and abiotic factors such as foraging behaviours in turbid conditions (Chamberlain & Ioannou, 2019), foraging frequency in the presence of boat noise (Pieniazek, Mickle, & Higgs, 2020) and habitat foraging occurrence and frequency with predation risk (Reinthal & Lewis, 1986). Currently, noninvasive methods to determine the rate of sea grass versus algal consumption by grazing fish such as luderick are rarely used. Instead, invasive methods including the capture and killing of herbivorous fish are often necessary to analyse stomach contents or perform isotope analysis to determine grazing rate and preferences (Raubenheimer, Zemke-White, Phillips, & Clements, 2005; Waltham & Connolly, 2006;

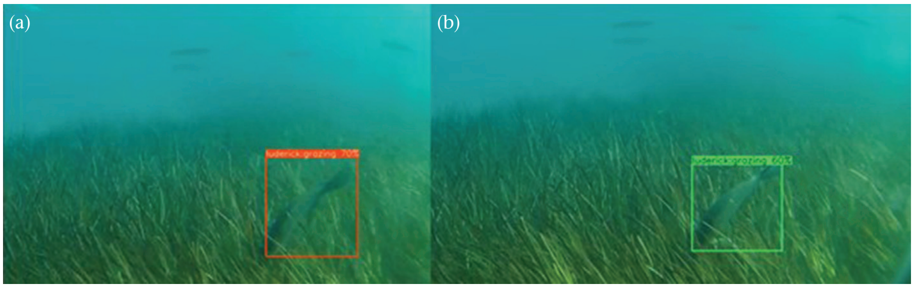

Figure 3. Two adjoining frames from a video of grazing behaviour displaying (a) an incorrect (false negative) and (b) a correct (true positive) assignment.

Dromard, Bouchon-Navaro, Harmelin-Vivien, & Bouchon, 2015). The potential application of our analytical tool could supplement or eventually, with further advances in computer vision, replace the need for invasive and lethal methods when coupled with other known variables such as the volume of algae removed per blade, fish size or total time grazing. There is a considerable upside for further research determining what other behaviours can be identified using this and other deep learning methods.

While it is possible to quantify behaviours using automation techniques, behaviours may not be effectively captured by in situ computer vision techniques due to camera limitations in environmental conditions such as increased water turbidity or decreased light availability (Gray et al., 2011, 2014; Van der Sluijs et al., 2011). However, technological advances could lead to deep learning techniques being effective on a range of other image-based data including sonar, which is not dependent on these environmental conditions (Enders, Danco, Podemski, & Wlasichuk, 2016). Using sonar in conjunction with deep learning techniques is possible for researchers to investigate and warrants further research into automated methods for other image-based data (Enders et al., 2016). Similar to the issue of capturing small or cryptic animals on camera, it is probable that researchers do not currently attempt to collect footage in environments that have poor visibility, and therefore cannot be manually analysed. Additionally, animals may frequent a number of different habitats or locations and behaviours may change depending on the environment. Ditria, Lopez-Marcano, Sievers, Jinks, and Connolly (2020b) showed that due to domain shift, where the training data are not an accurate enough representation of the real-world data they encounter, some models may not be transferable across habitat types. This may be a particular issue if behaviours differ markedly between habitats; for example, our target species exhibits epiphytic algal grazing in sea grass beds, but it exhibits a plucking motion when grazing on macroalgae in rock habitats. Further studies into how models perform when trained on multiple behaviours is needed.

Deep learning offers scientists a tool that is faster than manual analysis; however, there is still some subjectivity resulting from the initial categorization of behaviours (Han et al., 2018). What constitutes a specific behaviour is subject to interpretation by the person annotating the training data, as is the case in manual analysis of video footage. In our study, observing the visual overlay of computer-predicted behaviours on videos (see example in Fig. 3), we discerned that many of the grazing frames that are missed are those that are at the start of the grazing behaviours. This could be explained by the subjectivity of the manual annotator determining when the grazing behaviour begins. A difference in optical flow pixel movement, which the computer relies on to classify grazing behaviour, may not occur until one or several frames later.

Although research into combining optical flow and deep learning techniques has classified behaviours in the laboratory (e.g. Nguyen et al., 2019), the ability to automate the analysis of video data collected in the field presents a unique set of challenges that are rarely encountered in controlled systems. These may include varying distances of the subject from the camera, and environmental factors such as different weather conditions or varying water clarity, which can all affect model performance, especially if these conditions have not been encountered in training. Despite these environmental challenges, attempts to use deep learning for behavioural studies are beginning to be applied, especially in terrestrial environments. For example, Graving et al. (2019) classified a range of large, terrestrial animals displaying different poses and postures from still images using deep learning methods. We expect studies expanding the utility and versatility of deep learning algorithms to lead to a better understanding of how to analyse and automate the identification of animal behaviour.

We have shown that coupling deep learning with optical flow is a promising method to automate the analysis of grazing behaviours of a target species from underwater videos in unconstrained environments. The method seems ideally suited for questions about the frequency of specific behaviours, where the technique can be used to complement traditional techniques, reducing the time needed for manual analysis. Deep learning techniques can automate the analysis of underwater video data and obtain information on the distribution, abundance and behaviours of animals, signalling that this technology is likely to be important for the future of raw data analysis of images and videos to improve research efficiency.

Author Contributions

E.M.D. and R.M.C. conceptualized the study. E.M.D. did the field work and AI experiments. E.M.D. and E.L.J. developed the deep learning, interpreted model outputs and designed the training user interface. R.M.C. provided resources. E.M.D. led the writing with contributions from all authors.

Funding

RMC was supported by a Discovery Project from the Australian Research Council (DP180103124). All authors were supported by the Global Wetlands Project.

Declaration of Interest

The study was not carried out in the presence of any personal, professional or financial relationships that could potentially be construed as a conflict of interest.

Acknowledgments

We thank the volunteers who assisted in manual training of the deep learning algorithm. RMC was supported by a Discovery Project from the Australian Research Council (DP180103124). All authors were supported by the Global Wetlands Project.

Supplementary Material

Supplementary material associated with this article is available in the online version at https://doi.org/10.1016/j.anbehav.2021.04.018.

References

Altmann, J. (1974). Observational study of behavior: Sampling methods. *Behaviour, 49*(3–4), 227–266.

Browning, E., Bolton, M., Owen, E., Shoji, A., Guilford, T., & Freeman, R. (2018). Predicting animal behaviour using deep learning: GPS data alone accurately predict diving in seabirds. *Methods in Ecology and Evolution, 9*(3), 681–692. https://doi.org/10.1111/2041-210x.12926

Chamberlain, A. C., & Ioannou, C. C. (2019). Turbidity increases risk perception but constrains collective behaviour during foraging by fish shoals. *Animal Behaviour, 156*, 129–138.

Ditria, E. M., Lopez-Marcano, S., Sievers, M., Jinks, E. L., Brown, C. J., & Connolly, R. M. (2020a). Automating the analysis of fish abundance using object detection: Optimizing animal ecology with deep learning. *Frontiers in Marine Science, 7*, 429.

Ditria, E. M., Lopez-Marcano, S., Sievers, M., Jinks, E. L., & Connolly, R. M. (2020b). Deep learning for automated analysis of fish abundance: The benefits of training across multiple habitats. *Environmental Monitoring and Assessment*, (698), 192. https://doi.org/10.1007/s10661-020-08653-z

Dromard, C. R., Bouchon-Navaro, Y., Harmelin-Vivien, M., & Bouchon, C. (2015). Diversity of trophic niches among herbivorous fishes on a Caribbean reef (Guadeloupe, Lesser Antilles), evidenced by stable isotope and gut content analyses. *Journal of Sea Research, 95*, 124–131.

Enders, E. C., Danco, V., Podemski, C., & Wlasichuk, C. (2016). *Analysing the Impact of Freshwater Aquaculture on Wild Fish Populations Using Dual Frequency Identification Sonar (DIDSON) Technology*. Ottawa, Canada: Fisheries and Oceans Canada Ecosystems and Oceans Science Central and Arctic.

Espinoza, M., Farrugia, T. J., Webber, D. M., Smith, F., & Lowe, C. G. (2011). Testing a new acoustic telemetry technique to quantify long-term, fine-scale movements of aquatic animals. *Fisheries Research, 108*(2–3), 364–371.

Everingham, M., Van Gool, L., Williams, C. K., Winn, J., & Zisserman, A. (2010). The pascal visual object classes (voc) challenge. *International Journal of Computer Vision, 88*(2), 303–338.

Farnebäck, G. (2003). Two-frame motion estimation based on polynomial expansion. In J. Bigun, & T. Gustavsson (Eds.), *Image Analysis. SCIA 2003. Lecture Notes in Computer Science* (Vol. 2749, pp. 363–370). Berlin, Germany: Springer.

Golkarnarenji, G., Kouzani, A. Z., Semianiw, N. I., Goodall, D., Gilbert, D., & Driscoll, D. (2018). Automatic detection of moving Baw Baw frogs in camera trap videos. In *IEEE International Conference on Mechatronics and Automation (ICMA), Changchun, China, 2018* (pp. 1112–1116).

González-Rivero, M., Beijbom, O., Rodriguez-Ramirez, A., Bryant, D. E., Ganase, A., Gonzalez-Marrero, Y., et al. (2020). Monitoring of coral reefs using artificial intelligence: A feasible and cost-effective approach. *Remote Sensing, 12*(3), 489.

Graving, J. M., Chae, D., Naik, H., Li, L., Koger, B., Costelloe, B. R., et al. (2019). DeepPoseKit, a software toolkit for fast and robust animal pose estimation using deep learning. *eLife, 8*, Article e47994.

Gray, P. C., Bierlich, K. C., Mantell, S. A., Friedlaender, A. S., Goldbogen, J. A., & Johnston, D. W. (2019). Drones and convolutional neural networks facilitate automated and accurate cetacean species identification and photogrammetry. *Methods in Ecology and Evolution, 10*(9), 1490–1500.

Gray, S. M., Bieber, F. M. E., Mcdonnell, l. H., Chapman, l. J., & Mandrak, N. E. (2014). Experimental evidence for species-specific response to turbidity in imperilled fishes. *Aquatic Conservation: Marine and Freshwater Ecosystems, 24*(4), 546–560.

Gray, S. M., Sabbah, S., & Hawryshyn, C. W. (2011). Experimentally increased turbidity causes behavioural shifts in Lake Malawi cichlids. *Ecology of Freshwater Fish, 20*(4), 529–536.

Gray, P. C., Fleishman, A. B., Klein, D. J., McKown, M. W., Bézy, V. S., Lohmann, K. J., et al. (2019). A convolutional neural network for detecting sea turtles in drone imagery. *Methods in Ecology and Evolution, 10*(3), 345–355.

Gultekin, G. K., & Saranli, A. (2013). An FPGA based high performance optical flow hardware design for computer vision applications. *Microprocessors and Microsystems, 37*(3), 270–286.

Han, S., Taralova, E., Dupre, C., & Yuste, R. (2018). Comprehensive machine learning analysis of Hydra behavior reveals a stable basal behavioral repertoire. *eLife, 7*, Article e32605.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778).

Himawan, I., Towsey, M., Law, B., & Roe, P. (2018). Deep learning techniques for koala activity detection. In *INTERSPEECH* (pp. 2107–2111).

Hou, J., He, Y., Yang, H., Connor, T., Gao, J., Wang, Y., et al. (2020). Identification of animal individuals using deep learning: A case study of giant panda. *Biological Conservation, 242*, 108414.

Hughey, L. F., Hein, A. M., Strandburg-Peshkin, A., & Jensen, F. H. (2018). Challenges and solutions for studying collective animal behaviour in the wild. *Philosophical Transactions of the Royal Society B: Biological Sciences, 373*(1746), 20170005.

Jovanović, V., Svendsen, E., Risojević, V., & Babić, Z. (2018). Splash detection in fish Plants surveillance videos using deep learning. In *2018 14th Symposium on Neural Networks and Applications (NEUREL)* (pp. 1–5).

Konovalov, D. A., Hillcoat, S., Williams, G., Birtles, R. A., Gardiner, N., & Curnock, M. I. (2018). Individual minke whale recognition using deep learning convolutional neural networks. *Journal of Geoscience and Environment Protection, 6*, 25–36.

Ladds, M. A., Thompson, A. P., Kadar, J.-P., Slip, D. J., Hocking, D. P., & Harcourt, R. G. (2017). Super machine learning: Improving accuracy and reducing variance of behaviour classification from accelerometry. *Animal Biotelemetry, 5*(1), 8.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436–444.

Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3431–3440).

Nguyen, N. G., Phan, D., Lumbanraja, F. R., Faisal, M. R., Abapihi, B., Purnama, B., et al. (2019). Applying deep learning models to mouse behavior recognition. *Journal of Biomedical Science and Engineering, 12*(2), 183–196.

Norouzzadeh, M. S., Nguyen, A., Kosmala, M., Swanson, A., Palmer, M. S., Packer, C., et al. (2018). Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning. *Proceedings of the National Academy of Sciences, 115*(25), E5716–E5725.

Pieniazek, R. H., Mickle, M. F., & Higgs, D. M. (2020). Comparative analysis of noise effects on wild and captive freshwater fish behaviour. *Animal Behaviour, 168*, 129–135.

Raubenheimer, D., Zemke-White, W., Phillips, R., & Clements, K. (2005). Algal macronutrients and food selection by the omnivorous marine fish *Girella tricuspidata*. *Ecology, 86*(10), 2601–2610.

Reinthal, P. N., & Lewis, S. M. (1986). Social behaviour, foraging efficiency and habitat utilization in a group of tropical herbivorous fish. *Animal Behaviour, 34*(6), 1687–1693.

Rosenthal, M. F., Gertler, M., Hamilton, A. D., Prasad, S., & Andrade, M. C. (2017). Taxonomic bias in animal behaviour publications. *Animal Behaviour, 127*, 83–89.

Strout, J., Rogan, B., Seyednezhad, S. M., Smart, K., Bush, M., & Ribeiro, E. (2017). Anuran call classification with deep learning. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 2662–2665).

Sun, X., Shi, J., Liu, L., Dong, J., Plant, C., Wang, X., et al. (2018). Transferring deep knowledge for object recognition in Low-quality underwater videos. *Neurocomputing, 275*, 897–908. https://doi.org/10.1016/j.neucom.2017.09.044

Torney, C. J., Lloyd-Jones, D. J., Chevallier, M., Moyer, D. C., Maliti, H. T., Mwita, M., et al. (2019). A comparison of deep learning and citizen science techniques for counting wildlife in aerial survey images. *Methods in Ecology and Evolution, 10*(6), 779–787.

Valletta, J. J., Torney, C., Kings, M., Thornton, A., & Madden, J. (2017). Applications of machine learning in animal behaviour studies. *Animal Behaviour, 124*, 203–220.

Van der Sluijs, I., Gray, S. M., Amorim, M. C. P., Barber, I., Candolin, U., Hendry, A. P., et al. (2011). Communication in troubled waters: Responses of fish communication systems to changing environments. *Evolutionary Ecology, 25*(3), 623–640.

Villon, S., Mouillot, D., Chaumont, M., Darling, E. S., Subsol, G., Claverie, T., et al. (2018). A Deep learning method for accurate and fast identification of coral reef fishes in underwater images. *Ecological Informatics, 48*, 238–244.

Villon, S., Mouillot, D., Chaumont, M., Subsol, G., Claverie, T., & Villéger, S. (2020). A new method to control error rates in automated species identification with deep learning algorithms. *Scientific Reports, 10*(1), 1–13.

Walker, J., Gupta, A., & Hebert, M. (2015). Dense optical flow prediction from a static image. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 2443–2451).

Waltham, N. J., & Connolly, R. M. (2006). Trophic strategies of garfish, Arrhamphus sclerolepis, in natural coastal wetlands and artificial urban waterways. *Marine Biology, 148*(5), 1135–1141.

Weinstein, B. G. (2017). A computer vision for animal ecology. *Journal of Animal Ecology, 87*(3), 533–545.

Xie, J., & Zhu, M. (2019). Handcrafted features and late fusion with deep learning for bird sound classification. *Ecological Informatics, 52*, 74–81.

Xu, L., Bennamoun, M., An, S., Sohel, F., & Boussaid, F. (2019). Deep learning for marine species recognition. In V. Balas, S. Roy, D. Sharma, & P. Samui (Eds.), *Advances in computational intelligence* (pp. 129–145). Berlin, Germany: Springer.

Xu, W., & Matzner, S. (2018). Underwater fish detection using deep learning for water power applications. In *2018 International Conference on Computational Science and Computational Intelligence (CSCI)* (pp. 313–318).

Yang, A., Huang, H., Zhu, X., Yang, X., Chen, P., Li, S., et al. (2018). Automatic recognition of sow nursing behaviour using deep learning-based segmentation and spatial and temporal features. *Biosystems Engineering, 175*, 133–145.

Zhou, C., Xu, D., Chen, L., Zhang, S., Sun, C., Yang, X., et al. (2019). Evaluation of fish feeding intensity in aquaculture using a convolutional neural network and machine vision. *Aquaculture, 507*, 457–465.

Appendix

Table A1
The average performance of the three models per video with spatiotemporal filtering

Video	TP	FP	FN	R%	P%	F1%
1	18.0	14.3	0.0	100.0	55.7	71.5
2	58.7	29.3	10.3	85.0	66.7	74.7
3	0.3	0.0	7.7	4.2	33.3	7.4
4	38.7	0.0	17.3	69.0	100.0	81.5
5	10.3	0.0	0.7	93.9	100.0	96.8
6	49.0	0.0	1.0	98.0	100.0	99.0
7	3.7	0.0	6.3	36.7	100.0	52.4
8	10.3	0.0	7.7	57.4	100.0	72.8
9	19.0	0.0	1.0	95.0	100.0	97.4
10	14.7	0.0	10.3	58.7	100.0	73.0
11	8.3	0.0	0.7	92.6	100.0	95.8
12	25.0	0.0	3.0	89.3	100.0	94.3
13	14.0	0.0	0.0	100.0	100.0	100.0
14	19.0	0.0	1.0	95.0	100.0	97.4
15	10.7	0.0	0.3	97.0	100.0	98.4
16	10.0	0.0	0.0	100.0	100.0	100.0
17	16.7	0.0	0.3	98.0	100.0	99.0
18	29.0	0.0	0.0	100.0	100.0	100.0
Overall	355.3	43.7	67.7	84.0	89.1	86.5

These metrics reflect the performance of the model at identifying the grazing behaviour of each annotated luderick frame from the video, not whether the individual behaviour event was detected or not. Overall recall (R), precision (P) and F1 scores are calculated based on overall true positives (TP), false positives (FP) and false negatives (FN) as opposed to representing a mean score across the 18 videos.